Cureus Journal of Computer Science

Open Access Original Article

Early-Stage Diabetes Mellitus Risk Prediction Using Light Gradient-Boosting Machine With Synthetic Minority Oversampling Technique Combined With Edited Nearest Neighbors on a Novel Southern Indian Clinical Dataset

Received 05/11/2025
Review began 06/03/2025
Review ended 08/03/2025
Published 09/01/2025

DOI:
https://doi.org/10.7759/s44389-025-06580-z

Pooja Kallappagol [1] , Sheetalrani R. Kawale [2] , Sharankumar Holaychi [3]

1. Department of Computer Science, Karnataka State Women's University, Bijapur, IND 2. Department of Computer Science, Karnataka State Akkamahadevi Women University, Vijayapura, IND 3. Department of Community Medicine, Koppal Institute of Medical Sciences, Koppal, IND

Corresponding author: Pooja Kallappagol, poojakallappagol30@gmail.com

Abstract

India has the world's second-highest diabetes rate. According to the most recent survey, 25 million people are at risk of getting type 2 diabetes, while 77 million are already suffering from it. Diabetes is one of the leading chronic diseases in India and around the world. Early diagnosis is essential for effective disease management and prevention of complications. However, current diagnostic methods fail to detect early-stage risk factors. This research introduces a region-specific, clinically validated dataset and proposes a Light Gradient-Boosting Machine (LightGBM)-based framework with Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors (SMOTE-ENN) for early-stage diabetes risk prediction. We proposed a novel Southern India Diabetes Dataset by curating clinical records from the Koppal Institute of Medical Sciences in Koppal and Dhanavantari Multispeciality Hospital, Jamkhandi, Karnataka, India, encompassing 1,185 patient data annotated with 17 clinical and demographic characteristics. To balance the dataset's intrinsic class imbalance, the SMOTE-ENN approach was employed to improve model generalization. Six models, such as Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, and Multilayer Perceptron, were evaluated, followed by a comparative performance analysis. Among these, LightGBM gives better performance, achieving a train accuracy of 97.09%, test accuracy of 97%, and area under the receiver operating characteristic curve of 0.99. These experimental results emphasize the potential of integrating hybrid resembling techniques with gradient boosting frameworks for accurate and clinically relevant early-stage diabetes risk prediction.

Categories: AI applications, Health Informatics, Machine Learning (ML)
Keywords: diabetes mellitus, machine learning, lightgbm, smote-enn, feature selection

Introduction

India has the second-highest population of individuals living with diabetes globally. The latest survey indicates that there are 77 million individuals suffering from type 2 diabetes, while an additional 25 million people might potentially get the condition [1]. This disease ranks among the earliest diseases documented in human history. It was initially referenced in an Egyptian document around 3,000 years ago. The term diabetes mellitus was first introduced by the Greek physician Aretaeus. The origin of the term diabetes comes from the Greek, signifying "to pass through," and the term mellitus is derived from Latin, referring to honey, which indicates sweetness. The annual mortality rate for diabetes surpasses that of HIV/AIDS, with an estimated death happening every 10 seconds. This significantly impacts the increase in poor health and early death [2]. Diabetes is a long-term condition characterized by high blood glucose levels due to the body's inadequate production, secretion, or absorption of the insulin hormone [3]. Diabetes is a condition marked by abnormally high levels of glucose in the blood, primarily due to either cellular insensitivity to insulin or a lack of insulin production [4]. Type 2 diabetes mellitus, commonly referred to as age-onset or adult-onset diabetes, represents the most widespread type of diabetes. This form of diabetes is generally managed well with changes in diet and the use of oral medications, as it is marked by a gradual onset that can persist for multiple years. The exact origins of diabetes mellitus remain unclear. Nonetheless, the results appear to be shaped by both genetic and environmental influences [5]. In its advanced stage, type 2 diabetes mellitus represents almost 90% of all diabetes diagnoses [6]. In affluent nations, type 2 diabetes represents 85-95% of all diabetes cases. The incidence of type 2 diabetes is on the rise is more pronounced in low- and middle-income nations, attributed to socio-cultural changes, urban development, aging populations, reduced physical activity, and sedentary habits [7-8]. In 2011, an estimated 366 million individuals were living with diabetes mellitus; projections indicate that this number will increase to 552 million by 2030 [9]. In 2019, approximately 77 million individuals in India received a diagnosis of diabetes, positioning the country as the second most populous nation affected by this condition worldwide [10]. The International Diabetes Foundation reports that there

How to cite this article
Kallappagol P, Kawale S R, Holaychi S (September 01, 2025) Early-Stage Diabetes Mellitus Risk Prediction Using Light Gradient-Boosting Machine With Synthetic Minority Oversampling Technique Combined With Edited Nearest Neighbors on a Novel Southern Indian Clinical Dataset. Cureus J Comput Sci 2 : es44389-025-06580-z. DOI https://doi.org/10.7759/s44389-025-06580-z

are currently number of adults dealing with diabetes worldwide stands at approximately 463 million, with projections indicating a rise to 700 million by the year 2045 [11]. While the other 10% is mostly ascribed to type 1 diabetes and gestational diabetes, type 2 diabetes accounts for 90% of all people diagnosed with diabetes. Some 374 million persons are at increasing risk of acquiring type 2 diabetes. Worldwide, this disease has claimed 4.2 million lives. WHO projections show that diabetes mellitus is the tenth most common cause of death globally [12]. Diabetes is a complicated disease whose fundamental etiology is still not completely known. A number of variables, including heredity, genetics, obesity, high cholesterol levels, too much oil consumption, a high-carbohydrate diet, sugar consumption, nutritional deficits, lack of exercise, overeating, stress, hypertension, and infectious diseases, could all affect the evolution of diabetes [13]. The three primary approaches for doing blood or urine tests meant to identify diabetes are the fasting plasma glucose test, the oral glucose tolerance test, and the glycated hemoglobin test (A1C). The application of traditional diabetes screening methods is limited in low-income nations because of their challenging and costly characteristics. In 2019, 77.0 million individuals received a diagnosis of diabetes, yet over half of the population continued to be undiagnosed. Recognizing the importance of identification is crucial, as early discovery can significantly mitigate the likelihood of severe consequences [14]. The implementation of automated diabetes prediction systems could effectively address this challenge. The prediction of diabetes can be achieved by utilizing machine learning techniques on existing data. These methodologies demand little effort and result in minimal expenses for forecasting. These procedures enable individuals to identify diabetes independently, without needing a physician's guidance. These algorithms reduce the time required for diagnosing illnesses and analyzing symptoms. The timely identification of diabetes is essential in modern diabetes epidemiology, as it can result in more serious long-term outcomes [15]. It is essential to anticipate diabetes within the community to facilitate the implementation of suitable treatments and preventive measures for its future occurrence. The scientific community has recently shifted its focus towards the early and accurate prediction of diabetes by employing sophisticated computational techniques. The manifestation of human concepts relies on the utilization of artificial intelligence and soft computing techniques. These systems support medical diagnosis, along with various applications concerning human health. Machine learning is an emerging field of computational algorithms that utilize data from the external environment to mimic human intelligence [16-20]. A field within computer science involves the extraction of patterns from data to create an understanding of previously unrecognized inputs [21]. This study aims to develop and assess predictive algorithms for evaluating the risk of early diabetes.

Related work

This section examines previous studies that used machine learning and deep learning techniques to predict early-onset diabetes. Kawale and Kallappagol [22] developed an AI-based system for the early prediction of diabetes, using deep learning models such as Convolutional Neural Network, Long Short-Term Memory, and Simple Recurrent Neural Network on the primary collected dataset. The Simple Recurrent Neural Network had exceptional classification performance, with the best accuracy of 99.99% among the assessed models. Sai et al. [23] developed an ensemble model that integrates Light Gradient-Boosting Machine (LightGBM), K-Nearest Neighbors, and AdaBoost to predict type 2 diabetes using the PIMA Indian Diabetes Dataset. The ensemble attained an accuracy of 90.76%, demonstrating that the integration of LightGBM with other algorithms may improve predictive outcomes, despite the foundational model exhibiting only modest accuracy. Jaiswal and Gupta [24] used an ensemble of CATBoost, LightGBM, and XGBoost. The outcomes were assessed using the F1 measure, sensitivity, and accuracy rate. The LightGBM model has a mean squared error of 0.04 and an accuracy of 96%. Their findings underscore the therapeutic significance of early detection in high-risk groups via the use of advanced deep learning techniques. Xue et al. [25] performed an extensive assessment of several machine learning models using the UCI Diabetes Dataset and concluded that LightGBM attained an accuracy of 88.46%. Their findings highlight LightGBM's capability in efficiently handling structured health data for early diagnosis, notwithstanding the greater efficacy of other models like Support Vector Machine. Şen et al. [26] used decision tree-based ensemble techniques (Bagging, Random Forest, and Extra Trees) and determined that the Extra Trees classifier attained a remarkable accuracy of 99.2%. The study demonstrated that ensemble learning improves the precision of predictions for early-stage diabetes diagnosis. The identified research gap from past studies is as follows: Earlier research has mostly used standard datasets like PIMA and UCI, instead of creating new benchmark datasets that effectively represent actual clinical scenarios or the demography of certain regions. The researchers' feature selection approaches and prediction models were not assessed by doctors, hence raising concerns over their medical accuracy and applicability. Third, the models' predictions were inadequately compared with actual outcomes, predicted values, and confidence level, which are crucial for validating the model's reliability and applicability in real healthcare contexts.

Materials And Methods

Proposed method

In our prior research, we used the 806 Southern India Diabetes Dataset (SIDD) dataset. In this study, we have increased the number of samples in the dataset to 1,185. This research presents a clinically driven, machine learning methodology for early-stage diabetes risk prediction. The proposed technique addresses

significant deficiencies in prior research by using real-world clinical data, assessing feature significance with clinicians, and validating predictions with confidence levels. The dataset has 1,185 samples annotated with 17 clinical and demographic attributes, including Age, Gender, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visual Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiffness, Alopecia, Obesity, and Class. In the phase of preprocessing, class imbalance is addressed using Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors (SMOTE-ENN) approach, which consistently synthesizes minority class samples. The Random Forest method is used to determine the most essential clinical markers associated with diabetes risk. Medical practitioners analyze and validate these qualities to ensure they are clinically relevant. The top 10 features identified through the Random Forest feature selection technique were used for model training. Model performance is measured using parameters such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Furthermore, predictions are confirmed on the test set through contrasting actual against expected results, as well as the related confidence levels, resulting in a transparent and reliable diagnostic tool for early diabetes identification. The proposed framework is illustrated in Figure *1*.

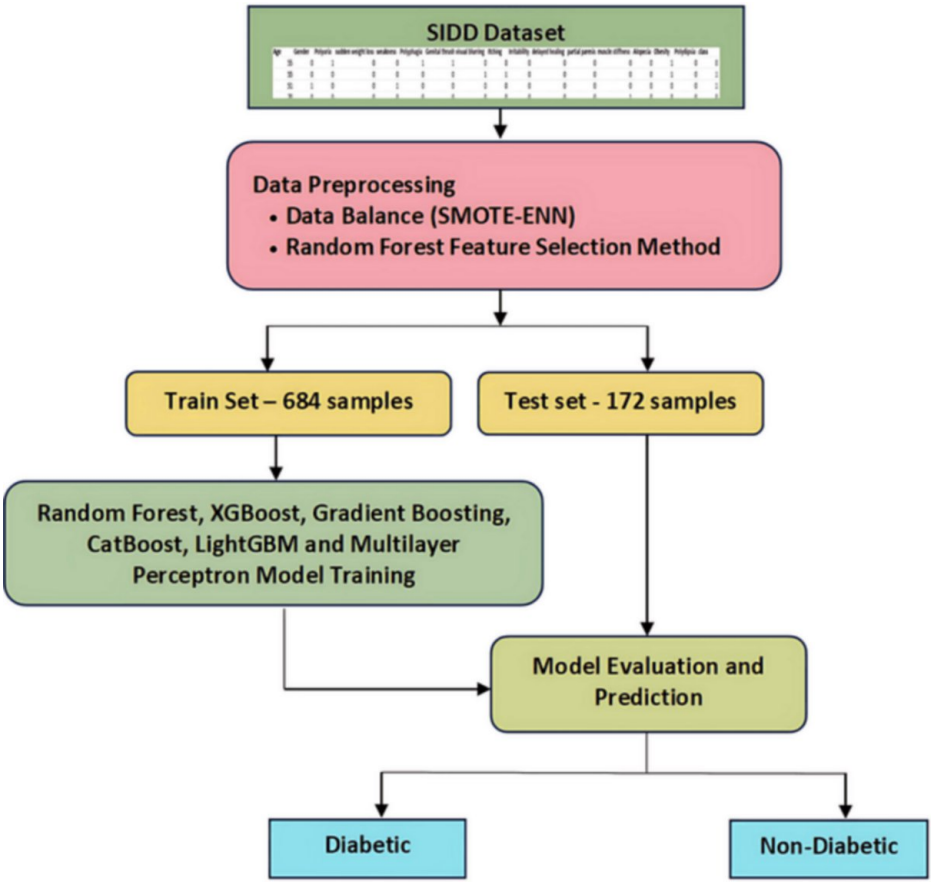

FIGURE 1: Conceptual Pipeline Process for the Suggested Framework

SIDD, Southern India Diabetes Dataset; SMOTE, Synthetic Minority Oversampling Technique; ENN, Edited Nearest Neighbors

Primary data collection

To propose an efficient prediction system, a novel dataset, referred to as the SIDD, was compiled from the Koppal Institute of Medical Sciences (KIMS), Koppal, and the Multi Specialist Hospital, Jamkhandi, Karnataka, India. The data collection process was conducted in accordance with ethical clearance guidelines to ensure patient confidentiality and data privacy. Patient information was obtained through a structured clinical assessment, wherein the presence of a symptom was recorded as “yes” and its absence as “no” based on blood test findings classified as diabetes or non-diabetes. The dataset consists of 1,185 patient records, each annotated with 17 clinical and demographic attributes: Age, Gender, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visual Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiffness, Alopecia, Obesity, and Class. Feature selection was based on established clinical indicators of early-stage diabetes mellitus and aligned with the feature set used in the Early-Stage Diabetes Risk Prediction dataset [27]. The UCI Early-Stage Diabetes

Risk Prediction Dataset has shown the potential of symptom-based predictive modeling and has been validated in its regional context, which may limit its applicability across diverse populations. To address this, the SIDD dataset ensures geographical relevance to the Southern Indian population.

Dataset description

The dataset has 1,185 patient records, each annotated with 17 clinical and demographic characteristics. The features include age, gender, and symptomatology such as Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visual Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiffness, Alopecia, and Obesity. The SIDD dataset comprises around 70.56% of records categorized as diabetes and 29.44% as non-diabetic. The description of the dataset is shown in Table *1*.

Sl. No.	Features	Description
1	Age	Age of the individual (range: 14–85 years)
2	Gender	Biological sex (Male/Female)
3	Polyuria	Excessive urination (Yes/No)
4	Sudden Weight Loss	Unexplained weight loss (Yes/No)
5	Weakness	General body weakness (Yes/No)
6	Polyphagia	Excessive hunger (Yes/No)
7	Genital Thrush	Itching or burning sensation in the genital area (Yes/No)
8	Visual Blurring	Blurred vision (Yes/No)
9	Itching	Itching on skin or body (Yes/No)
10	Irritability	Increased irritability (Yes/No)
11	Delayed Healing	Slow healing of wounds (Yes/No)
12	Partial Paresis	Partial paralysis or muscle weakness (Yes/No)
13	Muscle Stiffness	Stiffness in muscles (Yes/No)
14	Alopecia	Hair loss or balding (Yes/No)
15	Obesity	Presence of obesity (Yes/No)
16	Polydipsia	Excessive thirst (Yes/No)
17	Class	Diabetes or Non-diabetes

TABLE 1: Description of Clinical and Demographic Features Included in the Dataset

Handling class imbalance using SMOTE-ENN

This study used a hybrid resampling method known as SMOTE-ENN to mitigate the inherent class imbalance in the dataset. The SMOTE-ENN approach was proposed as an enhancement to traditional over sampling and insufficient sampling techniques, integrating their respective advantages to provide superior data balance. SMOTE method is first used to generate synthetic representations for the minority class. Synthetic samples are generated by interpolating among the feature-space neighbors of minority instances, hence augmenting minority representation by the reproduction of existing data. This method alleviates bias resulting from imbalanced class distributions during model training. Subsequent to synthetic augmentation, ENN is used to refine the data. ENN operates by evaluating each instance against its K-nearest neighbors (often k = 3) and eliminating instances having class labels that differ from the predominant labels among its neighbors. This reduces the probability of distracting, overlapping, or borderline samples that may impair the model's prediction efficacy. The integration of SMOTE and ENN enhances training set accuracy by (i) augmenting the minority class and (ii) refining class borders via noise reduction. Recent studies indicate that SMOTE-ENN surpasses conventional balancing methods across several domains, enhancing classifier metrics such as accuracy, recall, and F1-score [28-29]. Consequently, this work used SMOTE-ENN to preprocess the training data before classifier development, aiming to enhance the robustness and generalizability of the prediction models.

Feature selection using random forest feature importance

The feature selection is a crucial phase in the development of machine learning models, especially in healthcare applications where datasets often include redundant or minimally predictive factors [30]. This work used Random Forest-based feature importance analysis to identify and prioritize the key characteristics for predicting early-stage diabetes risk. Random Forest, an ensemble learning technique introduced by Breiman [31], provides robust classification performance together with inherent assessments of feature importance. Feature significance is often assessed by the mean reduction in Gini impurity or the average decrease in accuracy when a feature is permuted during tree building [32]. This allows the model to assess the relative importance of each element in forecasting the target variable. Following the use of SMOTE-ENN to balance the SIDD dataset, a Random Forest classifier was trained, and feature importance scores were calculated. The features age, polyuria, visual blurring, delayed healing, abrupt weight loss, obesity, polydipsia, gender, weakness, and polyphagia were recognized as the primary indicators for predicting diabetes risk, aligning with established clinical understanding [33]. Choosing the most relevant characteristics reduces model complexity, enhances interpretability, and increases training efficiency.

LightGBM classifier

The LightGBM is a gradient boosting framework that employs tree-based learning techniques. It was designed for efficiency and distributed computing, including two primary innovations: gradient-based one-side sampling and exclusive feature bundling. The gradient-based one-side sampling enhances efficiency by eliminating many data examples with little gradients and concentrating on those with significant gradients that substantially influence the calculation of information gain. This approach allows LightGBM to accurately assess information gain while using a reduced portion of the dataset. Simultaneously, exclusive feature bundling reduces the feature space by clustering mutually exclusive features, hence enhancing computing efficiency. The term "Light" in LightGBM denotes its exceptional speed and efficiency. LightGBM enhances traditional Gradient Boosting Decision Tree techniques for training duration, memory efficiency, accuracy, and scalability for large datasets. It further accommodates parallel and GPU learning configurations. It has been often used in tasks such as ranking, classification, and other machine learning applications. LightGBM distinguishes itself by its leafwise tree growth method, in contrast to the level wise splitting used by other boosting algorithms. LightGBM provides enhanced loss minimization and, thus, greater prediction accuracy compared to level-wise approaches by expanding each leaf with the maximum loss reduction at each iteration. However, for smaller datasets, fast leafwise expansion may enhance model complexity and lead to overfitting [34-36]. To address this challenge for our mid-sized SIDD dataset (about 1,185 instances), the experimental configuration used the LightGBM model with 200 estimators and a learning rate of 0.05. A fixed random state of 42 was used to ensure result repeatability. Figures *2* and 3 illustrate the proposed LightGMB architecture and flowchart, respectively.

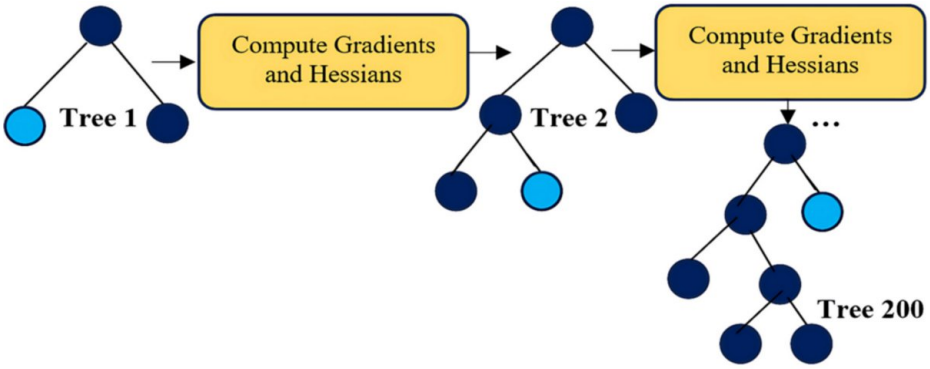

FIGURE 2: Proposed LightGBM Architecture

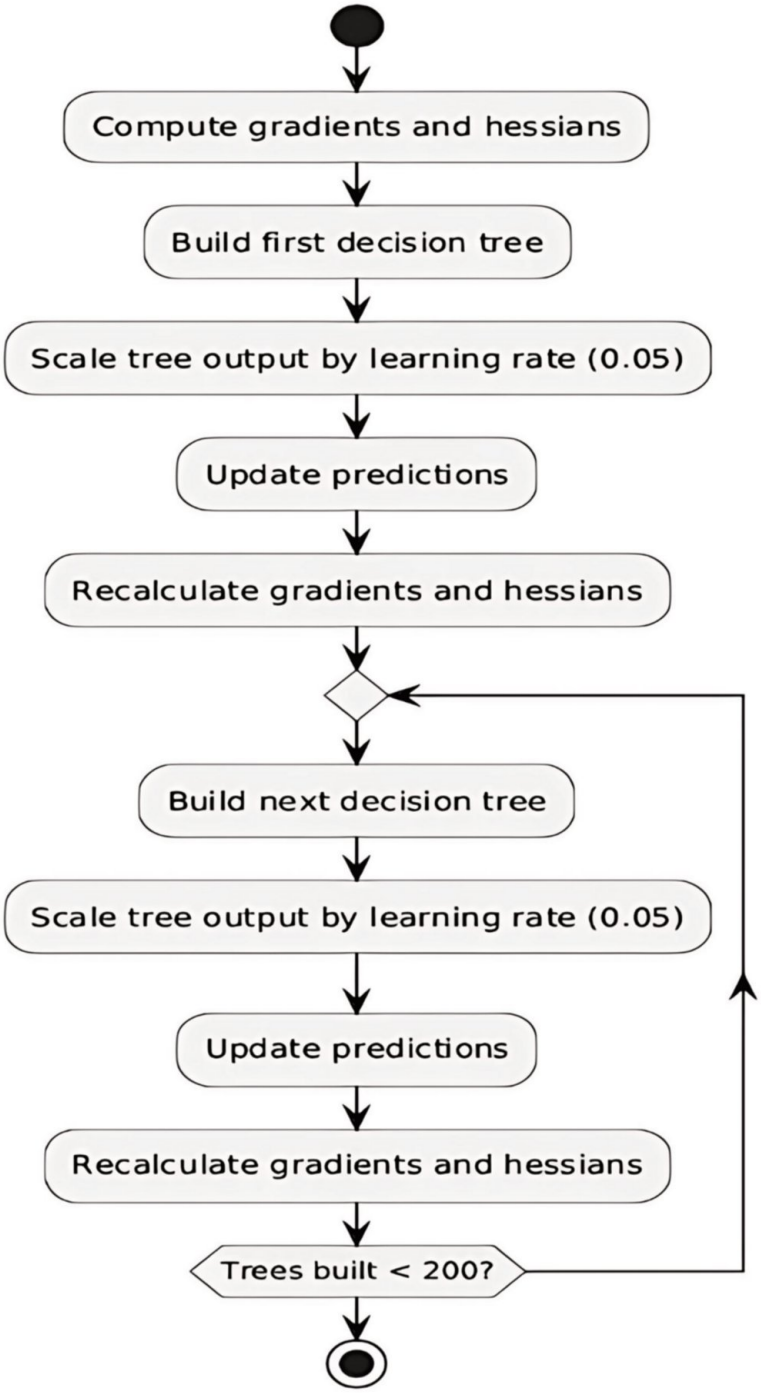

FIGURE 3: Flowchart for the LightGBM

Algorithm 1: Proposed LightGBM algorithm

Input:

· Training Dataset $D = \{(x_i, y_i)\}_{i=1}^{n})$

Parameters:

where:

D: training dataset

x_i: input features of the i^{th} sample

y_i: true label of the i^{th} sample

n: total number of samples

· Learning rate η (e.g., 0.05)

· Number of boosting iterations (trees) N = 200

Output:

· Final prediction model $F(x)$

Steps:

1. Initialize the Model:

Set the initial prediction $F_0(x)$ to minimize the loss function over all training samples.

2. For each boosting iteration $t = 1, 2, \ldots, N$, do:

a. Compute Gradients and Hessians:

For each training instance (x_i, y_i), compute:

Gradient:

$$g_i = \frac{\partial L\left(y_i, F_{t-1}\left(x_i\right)\right)}{\partial F_{t-1}\left(x_i\right)}$$

Parameters:

where:

y_i: true label

x_i: input features

$F_{t-1}(x_i)$: current model prediction

L: loss function

g_i: direction and size of error used to train the next weak learner

Hessian:

$$h_i = \frac{\partial^2 L\left(y_i, F_{t-1}\left(x_i\right)\right)}{\partial F_{t-1}\left(x_i\right)^2}$$

Parameters:

where:

y_i: true label

x_i: input features

$F_{t-1}(x_i)$: current prediction

L: loss function

h_i: Hessian (second derivative), captures curvature of loss

b. Build a Decision Tree:

Use (x_i, g_i, h_i) as input.

Grow tree leaf-wise: Split the leaf with maximum loss reduction.

Use histogram binning for split finding.

c. Scale Tree Output:
Multiply the output values at the tree's leaves by the learning rate η.

d. Update Model Predictions:

$$F_t(x) = F_{t-1}(x) + \eta \times \mathrm{Tree}_t(x)$$

Parameters:

where:

$F_{t-1}(x)$: previous prediction

η: learning rate

$\mathrm{Tree}_t(x)$: new tree fitted on gradients

e. (Optional Regularization):

Apply techniques like L1/L2 regularization on leaf outputs if configured.

3. End For

4. Return the Final Model $F(x) = F_N(x)$

Results

Experimental setup

The implementation of the early-stage diabetes risk prediction system was developed using Python version 3.11 on a Windows 11 operating environment, and used a Core i3 CPU with 16 GB of RAM. We constructed machine learning models based on Scikit-learn, LightGBM, and Imbalanced-learn.

The evaluation was performed using SIDD. Following the application of the SMOTE-ENN technique for class balancing, a Random Forest model was utilized to assess the significance of features within the dataset. The 10 features that garnered the highest relevance ratings were selected for subsequent model development. Six machine learning algorithms were trained and evaluated using the chosen features: Random Forest, XGBoost, CatBoost, Gradient Boosting, LightGBM, and Multilayer Perceptron (MLP). For final model selection, multiple evaluation criteria were considered; based on evaluation matrices and prediction confidence levels, XGBoost, LightGBM, and Catboost achieved a train accuracy of 97.09%, test accuracy of 97%, recall of 97%, precision and F1-score of 97%, and AUC-ROC of 0.99. However, LightGBM demonstrated slightly superior performance across other critical metrics. These differences can be attributed to the underlying architectural advantages of LightGBM. Specifically, LightGBM uses a leaf-wise tree growth strategy, which expands the leaf with the maximum loss reduction. This allows the model to capture complex patterns more effectively compared to XGBoost's level-wise (depth-wise) strategy or CatBoost's symmetric (oblivious tree) growth. Additionally, the dataset consisted of predominantly binary and numerical features, which LightGBM efficiently handled using its histogram-based split finding and gradient-based one-side sampling, allowing for faster convergence and lower training loss. Unlike CatBoost, which excels in datasets rich in high-cardinality categorical variables, our dataset's binary symptom structure favored LightGBM's architecture. These factors collectively explain LightGBM's slightly better performance and justify its selection as the final model.

The feature correlation matrix, showcasing the linear relationships among clinical and demographic data,

is presented in Figure 4. The significance of early identification is highlighted by the many robust positive correlations observed between polyuria, polydipsia, and the risk of diabetes. The low to moderate correlations found in most feature pairings indicate that the dataset has limited connections to the class. The matrix illustrates the independent predictive value of characteristics based on symptoms for model training.

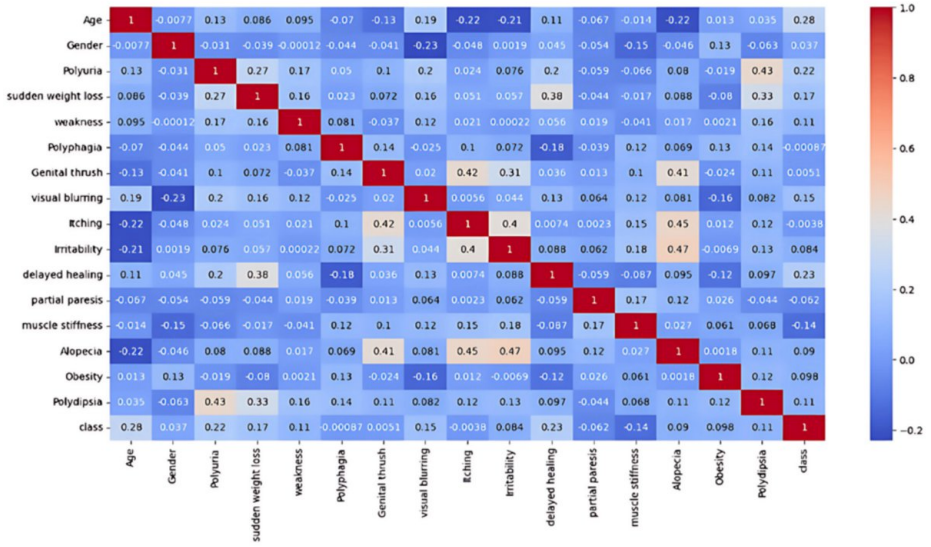

FIGURE 4: Features Correlation Matrix

The distribution of patients in the sample categorized by age is illustrated in Figure 5. The age groups of 50-60 and 60-70 represent the majority of instances, making up 27.3% and 31.8% of the dataset, respectively. The image presents the total count of diabetic cases (836) and non-diabetic cases (349), alongside the gender distribution, which comprises 584 male and 601 female participants, respectively. The demographics support the idea that the main groups affected by diabetes are the middle-aged and older individuals within the study population.

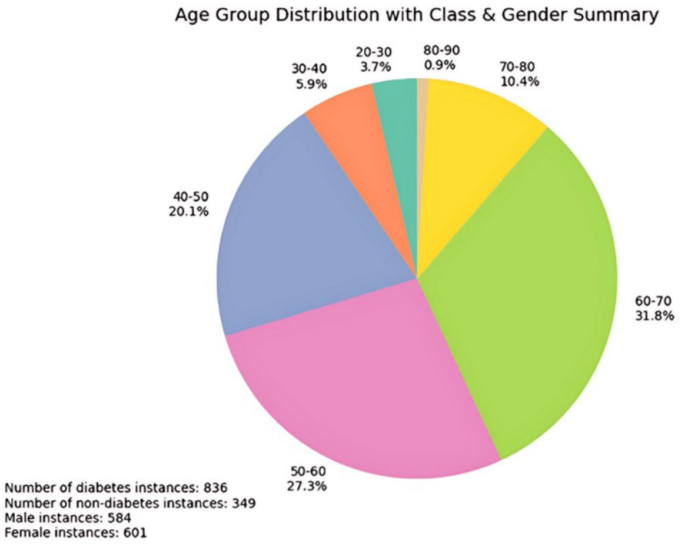

FIGURE 5: Diabetes Distribution According to the Ages

The distribution of patients categorized by the presence or absence of diabetes within the SIDD dataset is shown in Figure 6. The dataset consists of 70.5% individuals diagnosed with diabetes and 29.5% who are not diagnosed with the condition.

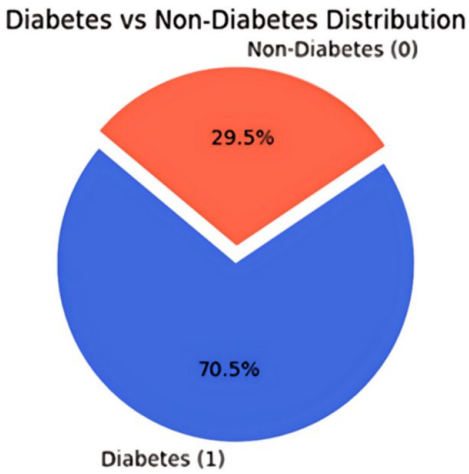

FIGURE 6: Class Distribution of Diabetic and Non-Diabetic in the Dataset

The symptoms polyuria, polydipsia, sudden weight loss, weakness, and visual blurring have a significantly higher number of positive instances, shown in Figure 7. This suggests a substantial correlation with the probability of developing diabetes in its early stages. Conversely, conditions such as vaginal thrush and partial paresis exhibit a reduced positive incidence rate. The age distribution indicates that the predominant segment of the patient population is middle-aged or senior.

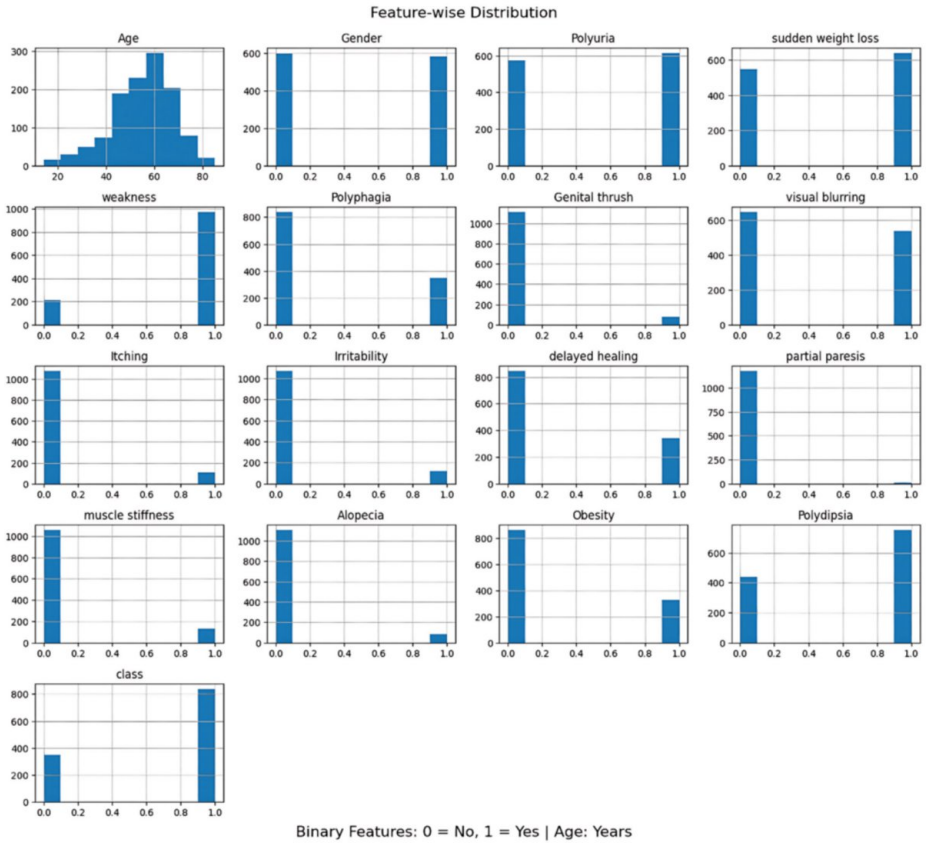

FIGURE 7: Distribution of Values Across Each Feature in the Dataset

The class distribution is shown in Table 2 both before and after the SMOTE-ENN resampling technique was applied. With 836 cases of diabetes and 349 cases of non-diabetes, the dataset initially displayed a moderate imbalance. Following SMOTE-ENN, there were 474 non-diabetic samples and 382 diabetic samples, bringing the groups into balance. A precise distribution between the two groups is ensured by the dataset's 684 training records and 172 testing records.

Class	Before SMOTE-ENN	After SMOTE-ENN	Train Set	Test Set
Diabetes	836	382	305	77
Non-Diabetes	349	474	379	95
Total	1,185	856	684	172

TABLE 2: Class-Wise Data Allocation Before and After SMOTE-ENN With Train–Test Split

SMOTE, Synthetic Minority Oversampling Technique; ENN, Edited Nearest Neighbors

The feature significance scores obtained from the Random Forest classifier are presented in Figure 8. Age, polyuria, visual blurring, delayed healing, sudden weight loss, obesity, polydipsia, gender, weakness, and polyphagia emerged as the top 10 risk factors for early-stage diabetes, with their relevance subsequently validated by domain experts. These 10 features exert the highest influence on model performance, while itching and partial paresis contribute minimally.

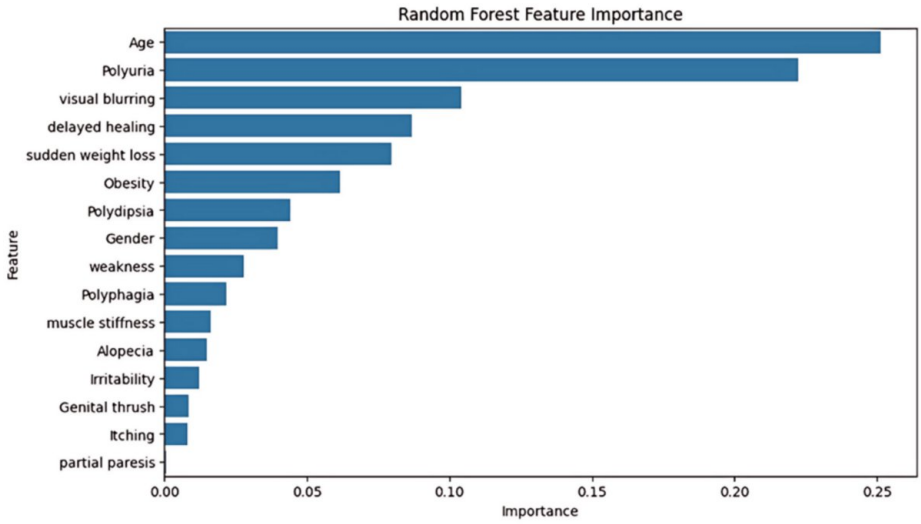

FIGURE 8: Random Forest Feature Importance

A comparison of the accuracy of various classifiers, such as Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, and MLP, utilizing two preprocessing methods SMOTE and SMOTE-ENN is shown in Figure 9. Models trained on datasets balanced with SMOTE-ENN regularly surpassed those trained on both original and SMOTE-balanced datasets. Notably, SMOTE-ENN achieved the greatest accuracy enhancements among all classifiers, demonstrating its effectiveness in mitigating class imbalance. This performance improvement underscores the significance of sufficient preprocessing in augmenting categorization outcomes.

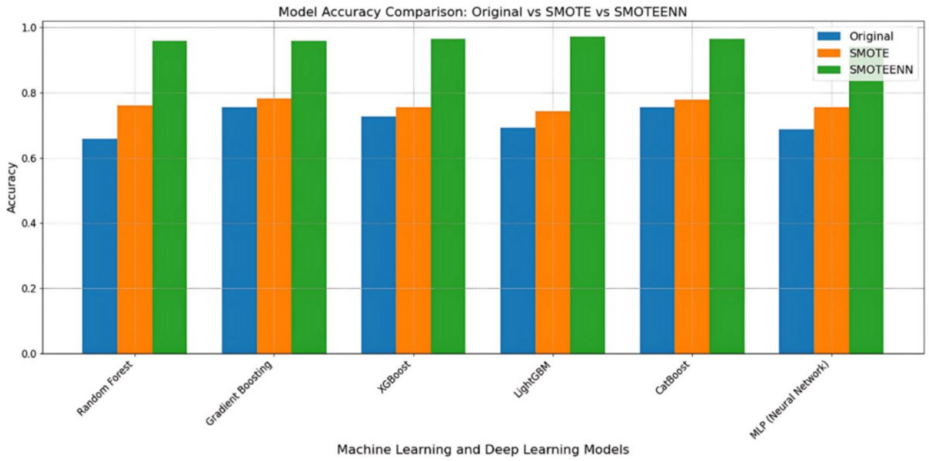

FIGURE 9: Model Accuracy Comparison With Original, SMOTE, and SMOTE-ENN

SMOTE, Synthetic Minority Oversampling Technique; ENN, Edited Nearest Neighbors

Table 3 presents the classification accuracies of six machine learning algorithms evaluated under three distinct sampling strategies: No Resampling, SMOTE, and SMOTE-ENN. Among these, SMOTE-ENN consistently demonstrated superior capability in mitigating class imbalance, yielding progressive accuracy improvements across all evaluated models. Notably, the LightGBM classifier achieved the highest performance, recording a training accuracy of 97.09% and a test accuracy of 97%. In comparison, XGBoost and CatBoost exhibited slightly lower training accuracies of 96.51% each, while maintaining an equivalent test accuracy of 97%.

Algorithms	Accuracy Without SMOTE and SMOTE-ENN	Accuracy With SMOTE	Accuracy With SMOTE-ENN	Test Accuracy
Random Forest	65.82%	76.12%	95.93%	96%
Gradient Boosting	75.53%	78.21%	95.93%	96%
XGBoost	72.57%	75.52%	96.51%	97%
LightGBM	69.20%	74.33%	97.09%	97%
CatBoost	75.53%	77.91%	96.51%	97%
Multi-Layer Perceptron	68.78%	75.52%	94.19%	94%

TABLE 3: Performance of the Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, and Multi-Layer Perceptron Classifiers

SMOTE, Synthetic Minority Oversampling Technique; ENN, Edited Nearest Neighbors

The class-wise precision, recall, and F1-scores that several classifiers achieved on the test set using SMOTE-ENN resampling are shown in Table 4. With precision and F1-score of 97%, LightGBM performed the best overall, followed by XGBoost and CatBoost, all of which had 96%. Interestingly, LightGBM demonstrated a better balance between recall and precision in both the diabetic and non-diabetic groups. With a macro-averaged F1-score of 94%, the MLP classifier, on the other hand, performed comparatively badly, indicating that ensemble-based methods were more appropriate for the early-stage diabetes risk prediction task.

Algorithms	Class	Precision	Recall	F1-score	Support
Random Forest	0	97%	96%	96%	95
	1	95%	96%	95%	77
	Macro avg	95%	95%	95%	172
	Weighted avg	95%	95%	95%	172
Gradient Boosting	0	96%	97%	96%	95
	1	96%	95%	95%	77
	Macro avg	96%	96%	96%	172
	Weighted avg	96%	96%	96%	172
XGBoost	0	96%	98%	97%	95
	1	97%	95%	96%	77
	Macro avg	97%	96%	96%	172
	Weighted avg	97%	97%	97%	172
LightGBM	0	97%	98%	97%	95
	1	97%	96%	97%	77
	Macro avg	97%	97%	97%	172
	Weighted avg	97%	97%	97%	172
CatBoost	0	97%	97%	97%	95
	1	96%	96%	96%	77
	Macro avg	96%	96%	96%	172
	Weighted avg	97%	97%	97%	172
Multi Layer Perceptron	0	97%	93%	95%	95
	1	91%	96%	94%	77
	Macro avg	94%	94%	94%	172
	Weighted avg	94%	94%	94%	172

TABLE 4: Classification Report Performance of Algorithms

The evaluation results of the LightGBM model on the SMOTE-ENN balanced dataset are shown in Figure *10*(a) and (b), which show the model's classification performance as shown by the confusion matrix and its discriminatory ability as indicated by the ROC curve. LightGBM exhibits exceptional prediction performance, properly classifying 93 out of 95 non-diabetic cases and 74 out of 77 diabetic cases with very few misclassifications, according to the confusion matrix (Figure *10*). Additionally, the model's outstanding sensitivity and specificity are validated by the ROC curve (Figure *10*), which displays a remarkable AUC value of 0.99.

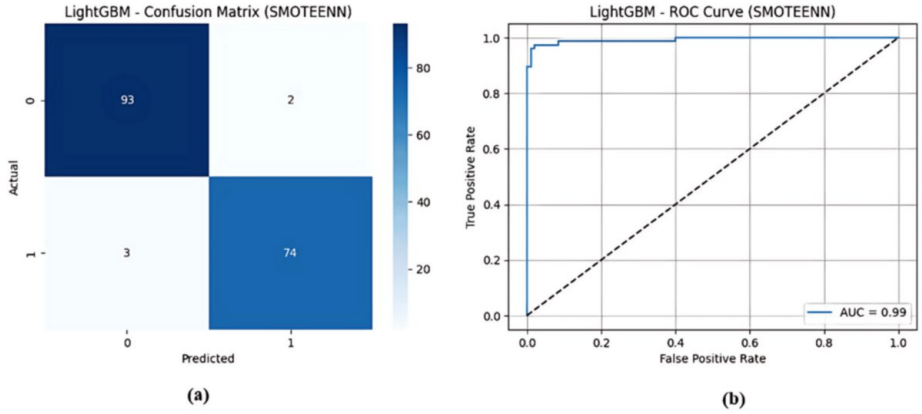

FIGURE 10: The Model's Classification Performance as Shown by (a) the Confusion Matrix and (b) the ROC Curve

AUC, Area Under The Curve; ROC, Receiver Operating Characteristic; SMOTE, Synthetic Minority Oversampling Technique; ENN, Edited Nearest Neighbors

The sample evaluation findings for the LightGBM model on the test set are shown in Table 5, including actual class, predicted class, and confidence levels. The model revealed excellent confidence levels for the majority of correct predictions, notably in positive diabetes patients. A few samples had poor confidence scores, indicating problematic borderline cases around the decision boundary. Overall, the LightGBM model performed well on the SMOTE-ENN balanced dataset, with high prediction certainty and accuracy.

Sl. No.	Actual Class	Predicted Class	Confidence Level
1	1	1	0.999923
2	1	1	0.999929
3	0	0	0.025638
4	1	1	0.991578
5	0	0	0.000148
6	1	1	0.999852
7	0	0	0.000890
8	1	1	0.999533
9	1	1	0.980290
10	0	0	0.001001

TABLE 5: Classification Results of LightGBM for the Test Set

Discussion

Experimental analysis

This work utilized a holistic experimental setup to investigate the performance of the proposed LightGBM-based model for predicting early-stage diabetic risk using the SIDD dataset. The dataset includes 1,185 physician-verified patient records with 17 clinical and demographic characteristics, which were acquired in accordance with the ethical norms. Considering the class imbalance (70.5% diabetic and 29.5% nondiabetic cases), the SMOTE-ENN approach was used to oversample the minority sample and under sample noisy samples, which can improve the robustness of training. Feature importance analysis was performed with Random Forest, which suggested polyuria, polydipsia, sudden weight loss, and visual blurring as the most predictive symptoms. These results agreed with established clinical experience, and were corroborated by clinicians. Six machine learning (Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, MLP) models were built and tested with the selected features. Of these,

LightGBM showed the best train and test accuracies as 97.09% and 97%, respectively, and had a test AUC of 0.99. The confusion matrix reveals a comparatively lower rate of misclassification, indicating robust predictive performance. Furthermore, the ROC curve corroborates the model's satisfactory discriminative ability, demonstrating its effectiveness in distinguishing between the target classes. Model performance measures, like precision, recall, and F1-score, also confirmed the superior performance of LightGBM, especially for the borderlines with high certainty. From the comparison results, SMOTE-ENN was demonstrated to consistently outperform other sampling approaches for all the models, suggesting the importance of hybrid resampling strategies for clinical datasets. These experimental results demonstrate the reliability and practical consequence of integrating SMOTE-ENN preprocessing and ensemble learning paradigms such as LightGBM to predict early-stage diabetes in regional populations with robust accuracy.

The performance of LightGBM-based models across various studies and datasets is examined in Table *6*. Srivatsan and Santhanam [37] obtained 91.84% on the Vanderbilt dataset, while Sai et al. [23] used an ensemble technique to score 90.76% on the PIMA dataset. Xue et al. [25] obtained an accuracy of 88.46% using LightGBM on the UCI dataset. The importance of clinically certified datasets was demonstrated by the significantly greater accuracy of 97.09% attained by the suggested strategy employing physician-verified SIDD-1185 data.

Authors	Methods	Dataset	Accuracy
Sai et al. [23]	LightGBM	PIMA Indian Diabetes Dataset	90.76%
Srivatsan and Santhanam [37]	LightGBM	Vanderbilt Biostatistics Diabetes Dataset	91.84 %
Xue et al. [25]	LightGBM	UCI Diabetes Dataset	88.46%
Proposed Method	LightGBM	SIDD-1185	97.09%

TABLE 6: Performance Comparison of the Proposed Method Across Different Datasets

SIDD, Southern India Diabetes Dataset; UCI, University of California, Irvine

Conclusions

The study developed a robust framework for early-stage diabetes risk prediction using a novel SIDD with 1,185 patient records. The proposed approach integrated SMOTE-ENN resampling, Random Forest feature selection, and LightGBM modeling. LightGBM achieved the highest training and test accuracies of 97.09% and 97%, respectively, along with an AUC of 0.99, thereby outperforming the other models tested. The model demonstrated strong precision, recall and F1-scores for both diabetic and non-diabetic classes. Results were validated by diabetes care specialists, enhancing clinical credibility and applicability. The clinically verified SIDD dataset and hybrid modeling approach led to improved performance compared to previous studies using standard datasets. Future work should focus on expanding the dataset diversity, exploring advanced feature selection for early indicators, and leveraging improved deep learning models. The framework shows potential to support early diabetes detection in real-world clinical settings, particularly for the Southern Indian population.

Appendices

The Southern India Diabetes Dataset (SIDD 1185) consists of 1,185 anonymized patient records, which were obtained from the Koppal Institute of Medical Sciences, Koppal, Karnataka, India, following a set of standard medical protocols for diagnosis. There are 17 clinical and demographic features that are related to the prediction task of early-onset diabetes in the dataset. The institutional ethics and patients' data protection guidelines were adhered and complied to during the data collection process. Dataset is not publicly available due to the sensitive nature of the data. However, it can be obtained from the corresponding author upon reasonable request. Requests for access may be submitted via the following link:

Dataset Access: https://drive.google.com/file/d/1DL9kocz1YvtX2Fk_Dn4dkUzbDUutAzSE/view?usp=sharing

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Pooja Kallappagol, Sheetalrani R. Kawale , Sharankumar Holaychi

Acquisition, analysis, or interpretation of data: Pooja Kallappagol, Sheetalrani R. Kawale , Sharankumar Holaychi

Drafting of the manuscript: Pooja Kallappagol, Sheetalrani R. Kawale , Sharankumar Holaychi

Critical review of the manuscript for important intellectual content: Pooja Kallappagol, Sheetalrani R. Kawale , Sharankumar Holaychi

Supervision: Sheetalrani R. Kawale

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Mohajan D, Mohajan HK: Historical view of diabetics mellitus: From ancient Egyptian polyuria to discovery of insulin. Studies in Social Science & Humanities. 2023, 2:26-34.
2. Raza A, Shabbir A, Shafique U, Ahmed N, Rafiq M: Investigation of diabetes mellitus transmission in humans by using time delay tool and numerical treatment approach. Modeling Earth Systems and Environment. 2025, 11:113. 10.1007/s40808-024-02274-y
3. Bereda G: Brief overview of diabetes mellitus. Diabetes Management. 2021, 11:21-27.
4. Ahmed S, Saeed S, Shubrook JH: Masqueraders: how to identify atypical diabetes in primary care. Journal of Osteopathic Medicine. 2021, 121:899-904. 10.1515/jom-2021-0129
5. Bereda G, Bereda G: The incidence and predictors of poor glycemic control among adults with type 2 diabetes mellitus in ambulatory clinic of Mettu Karl Referral Hospital, southwestern, Ethiopia: a prospective cross sectional study. Diabetes Updates. 2021, 7:1-6. 10.15761/du.1000155
6. Kengne AP, Ramachandran A: Feasibility of prevention of type 2 diabetes in low- and middle-income countries. Diabetologia. 2024, 67:763-772. 10.1007/s00125-023-06085-1
7. Liu J, Bai R, Chai Z, Cooper ME, Zimmet PZ, Zhang L: Low- and middle-income countries demonstrate rapid growth of type 2 diabetes: an analysis based on Global Burden of Disease 1990-2019 data. Diabetologia. 2022, 65:1339-1352. 10.1007/s00125-022-05713-6
8. Alam MM, Chowdhury SR, Azizul Islam SKM: Prevalence of chronic complications of type 2 diabetes mellitus: Results from a hospital OPD and personal chamber based cross sectional study. Southern Medical College Journal. 2022, 3:7-11.
9. Pradeepa R, Mohan V: Epidemiology of type 2 diabetes in India. Indian Journal of Ophthalmology. 2021, 69:2932-2938. 10.4103/ijo.ijo_1627_21
10. Teo ZL, Tham YC, Yu M, et al.: Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology. 2021, 128:1580-1591. 10.1016/j.ophtha.2021.04.027
11. Biswas T, Behera BK, Madhu NR: Technology in the management of type 1 and type 2 diabetes mellitus: Recent status and future prospects. Advances in Diabetes Research and Management. 2023, 111-136. 10.1007/978-981-19-0027-3_6
12. Ahmad E, Lim S, Lamptey R, Webb DR, Davies MJ: Type 2 diabetes. The Lancet. 2022, 400:1803-1820. 10.1016/s0140-6736(22)01655-5
13. Arokiasamy P, Salvi S, Selvamani Y: Global burden of diabetes mellitus. Handbook of Global Health. Haring R, Kickbusch I, Ganten D, Moeti M (ed): Springer, Cham; 2021. 1-44. 10.1007/978-3-030-05325-3_28-1
14. Tomic D, Shaw JE, Magliano DJ: The burden and risks of emerging complications of diabetes mellitus. Nature Reviews Endocrinology. 2022, 18:525-539. 10.1038/s41574-022-00690-7
15. Buja A, Caberlotto R, Pinato C, et al.: Health care service use and costs for a cohort of high-needs elderly diabetic patients. Primary Care Diabetes. 2021, 15:397-404. 10.1016/j.pcd.2020.12.002
16. Chung JKO, Xue H, Pang EWH, Tam DCC: Accuracy of fasting plasma glucose and hemoglobin A1c testing for the early detection of diabetes: A pilot study. Frontiers in Laboratory Medicine. 2017, 1:76-81. 10.1016/j.flm.2017.06.002
17. Drobek N, Sowa P, Jankowski P, et al.: Undiagnosed diabetes and prediabetes in patients with chronic coronary syndromes—An alarming public health issue. Journal of Clinical Medicine. 2021, 10:1981. 10.3390/jcm10091981
18. Alghamdi T: Prediction of diabetes complications using computational intelligence techniques. Applied Sciences. 2023, 13:3030. 10.3390/app13053030
19. Mujumdar A, Vaidehi V: Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019, 165:292-299. 10.1016/j.procs.2020.01.047
20. Ellahham S: Artificial intelligence: The future for diabetes care. The American Journal of Medicine. 2020, 133:895-900. 10.1016/j.amjmed.2020.03.033

21. Yamada S: First, do no harm: Critical appraisal of protein restriction for diabetic kidney disease. Diabetology. 2021, 2:51-64. 10.3390/diabetology2020005
22. Kawale SR, Kallappagol P: AI-driven neural networks for early-stage diabetes prediction. Journal of Computational Analysis & Applications. 2024, 33:740-751.
23. Sai MJ, Chettri P, Panigrahi R, Garg A, Bhoi AK, Barsocchi P: An ensemble of light gradient boosting machine and adaptive boosting for prediction of type-2 diabetes. International Journal of Computational Intelligence Systems. 2023, 16:14. 10.1007/s44196-023-00184-y
24. Jaiswal S, Gupta P: Ensemble approach: XGBoost, CATBoost, and LightGBM for diabetes mellitus risk prediction. 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India. 2022, 1-6. 10.1109/ICCSEA54677.2022.9936130
25. Xue J, Min F, Ma F: Research on diabetes prediction method based on machine learning. Journal of Physics: Conference Series. 2020, 1684:012062. 10.1088/1742-6596/1684/1/012062
26. Şen Ö, Keser SB, Keskin K: Early stage diabetes prediction using decision tree-based ensemble learning model. International Advanced Researches and Engineering Journal. 2023, 7:62-71. 10.35860/iarej.1188039
27. Early Stage Diabetes Risk Prediction. UCI Machine Learning Repository. (2020). https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset.
28. Batista GEAPA, Prati RC, Monard MC: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter. 2004, 6:20-29. 10.1145/1007730.1007735
29. Fernandez A, Garcia S, Herrera F, Chawla NV: SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research. 2018, 61:863-905. 10.1613/jair.1.11192
30. Guyon I, Elisseeff A: An introduction to variable and feature selection. Journal of Machine Learning Research. 2003, 3:1157-1182.
31. Breiman L: Random forests. Machine Learning. 2001, 45:5-32. 10.1023/a:1010933404324
32. Louppe G, Wehenkel L, Sutera A, Geurts P: Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems. 2013, 26:1-9.
33. ElSayed NA, Aleppo G, Aroda VR, et al.: Classification and diagnosis of diabetes: Standards of care in diabetes—2023. Diabetes Care. 2022, 46:S19-S40. 10.2337/dc23-s002
34. Ke G, Meng Q, Finley T, et al.: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017, 30:1-9.
35. Ke G, Meng Q, Finley T: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017, 30:1-9.
36. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G: Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports. 2020, 10:11981. 10.1038/s41598-020-68771-z
37. Srivatsan S, Santhanam T: Early onset detection of diabetes using feature selection and boosting techniques. ICTACT Journal on Soft Computing. 2021, 12:2474-2484. 10.21917/ijsc.2021.0354