Biomedical & Pharmacology Journal, June 2026. Vol. **19**(2), p. 1604-1616

An Ensemble Machine Learning Approach for Diabetes Classification and Prediction with Minimal Error: A Comprehensive Exploratory Data Analysis

Vishal Verma[1]*, Malik Shahzad Ahmed Iqbal[2], Satish Kumar[3], Vandna Rani Verma[4] and Alka Agrawal[1]

[1]Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, India.
[2]Department of Computer Science and Engineering, Acharya University, Karakul, Uzbekistan.
[3]Department of Computer Application, Integral University, Lucknow, India.
[4]Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India.
*Corresponding Author E-mail: vishalmgs93@gmail.com

https://dx.doi.org/10.13005/bpj/3443

(Received: 06 June 2025; accepted: 21 January 2026)

Diabetes is one of the chronic metabolic diseases and remains a serious public health problem worldwide. It results in multiple health problems, including high blood pressure, skin disorders, heart disease, kidney diseases, and eye damage. Many individuals with diabetes are undiagnosed for a long time. Early prediction is crucial in receiving timely treatment for the patient. The primary goal of this research is to construct an effective and reliable prediction model for early-stage diabetes, incorporating ensemble learning with novel data balancing techniques. In this study, researchers have utilized a Machine learning (ML) based ensemble classifier, Extra Trees Classifier (ETC), along with twelve data balancing techniques, namely SMOTE-ENN, Random Over Sampler, Instance Hardness Threshold, SMOTE Tomek, SMOTE, KMeans SMOTE, Borderline SMOTE, AllKNN, Random Under Sampler, Neighbourhood Cleaning Rule, Near Miss, and Tomek Links to predict diabetes. This study has utilized a newly extensive dataset, the Diabetes Prediction Dataset (DPD), taken from the Kaggle repository. The dataset consists of 100000 instances. The findings demonstrate that SMOTE-ENN combined with Extra Tree Classifier (SEETC) performs outstandingly. SEETC proposed model achieved 0.997, 0.993, 0.993, 0.997, 0.995, 0.995, 0.995, 0.3, and 0.7, respectively, in PrecisionNeg (PNeg), PrecisionPos (PPos), RecallNeg (RNeg), RecallPos (RPos), F1_ScoreNeg (F1Neg), F1_ScorePos (F1Pos), Accuracy (Acc), Type 1 Error Rate (T1E Rate) & Type 2 Error Rate (T2E Rate). The proposed model has been evaluated against other existing models, including Decision Tree (DT), Random Forest (RF), AdaBoost Classifier (ABC), and XGBoost (XGB). The results indicate that integrating SMOTE-ENN with ETC yields superior performance. This integration will be highly beneficial for diabetes prediction.

Keywords: Ensemble; Extra Tree Classifier; Healthcare; Imbalanced data; Machine learning; SMOTE-ENN.

Diabetes is a long-term condition that is rapidly increasing worldwide, posing one of the most significant health challenges in both developed and developing countries. According to the International Diabetes Federation (IDB) Diabetes Atlas 11th Edition 2025, around 589 million adults aged 20-79 globally are currently affected by diabetes. This figure is anticipated to rise to 643 million by 2030 and 853 million by 2050.[1] The IDB report highlights the rapid growth

of diabetes cases from 2000 to 2050. Diabetes occurs in humans when the body can't properly control blood sugar. The pancreas produces insulin and regulates blood sugar by enabling cells to absorb glucose for energy in the body. However, when insulin production and function become impaired, these cause an increase in blood sugar, which leads to diabetes. Generally, diabetes is classified as Type 1 (T1D), Type 2 (T2D), and Gestational diabetes. In T1D, the pancreas fails to produce insulin. While in T2D, the body either exhibits insulin resistance or fails to produce sufficient insulin to sustain normal glucose levels, and it primarily affects both young and adult populations. It is a chronic illness that affects how the body reacts to blood sugar. This leads to high blood glucose levels, and over time it can lead to serious health problems such as heart disease, neuropathy, kidney failure, and vision loss.[2] The risk of gestational diabetes is much higher while a woman is pregnant, particularly when blood sugar levels are not controlled. By doing so, women with gestational diabetes are going to be more exposed to the risk of developing T2D in their future lives.[3]

Early identification of T2D is crucial to reducing the risk of severe chronic disease and complications. This can be controlled through lifestyle changes, daily meditation, medical remedies, and ongoing monitoring. People may experience gradual damage to vital body organs until they are diagnosed with T2D late.[4] Traditional diagnostic procedures often involve testing an individual's glucose level, HbA1c (haemoglobin A1c), and fasting blood sugar. However, there is tremendous potential in using predictive ML algorithms to estimate the likelihood of diabetes using a variety of factors, which could provide a faster and more accessible way to screen large populations.[5]

In the past decade, ML approaches have emerged as highly effective methods for predicting diabetes. These approaches extract useful information by analyzing complex datasets by identifying trends and patterns that make it easier for researchers to build diagnostic models.[6] Accurate classification is the most crucial concern that helps clinicians as well as individuals with proper treatment planning procedures.[7] Though various researchers have proposed individual and ensemble models but these models face several limitations, including accuracy issues, high error rates, and overfitting etc. Hence, further researches need to overcome these limitations and needs to come up with various models aiming to enhance the accuracy, reduce the high error rate, and prevent overfitting. Therefore, the researchers have tried to address these limitations by aiming to increase the accuracy of diagnosis and to minimize the false-positive and false-negative error rates.

To address this gap, the authors have conducted an empirical study into how well the Extra Tree Classifier (ETC) performs under twelve resampling techniques. ETC is an ensemble method that has the potential to mitigate the problem of imbalanced data overfitting.[8] It is also known as an Extremely Randomized Trees Classifier, which uses randomness to create multiple Decision Trees and then aggregates their predictions for the classification. This helps to create a more robust classifier and reduces the chances of overfitting. It's capable of handling large and complex datasets well.[9]

In addition, a comprehensive Exploratory Data Analysis (EDA) has been carried out to facilitate the understanding of the dataset and to assist in achieving the goal. EDA is critical to reveal patterns, relationships, and possible anomalies in the dataset.[10] Accordingly, the researchers have used EDA to acquire insights from the dataset. Furthermore, EDA and ML provide a comprehensive method for enhancing the precision and dependability of diabetes prediction, with the potential for a significant impact in healthcare applications.[11] During EDA, the two features, gender and smoking history, were noted to suffer from significant data quality issues. For example, many records did not have gender information, and about 34% of cases had missing values in the smoking history attribute. To keep the dataset intact and good quality, records with no gender were dropped, which will not affect the result. The column about a smoking history was also removed from the available dataset to avoid the possibility of bias and to ensure the robustness of the model. Thus, the aim of the paper can be summarized as building a better diabetes prediction model with an error rate as low as possible. The clear research objectives are structured as:

• To eliminate bias, a comprehensive Exploratory Data Analysis (EDA) on the diabetes dataset has been performed.
• To find out the best resampling technique with the lowest error rate, twelve resampling techniques have been investigated for their impact in handling class imbalance within the dataset.
• To evaluate the efficiency of the proposed SEETC model on various metrics.
• To compare the performance of the proposed model with existing ML models to establish its effectiveness and superiority.

This study article is organized as follows: Section 2 reviews related work on the usage of various machine learning models and data resampling techniques for diabetes prediction. The methodology is presented in Section 3. The experimental study results are highlighted and discussed in Section 4. Finally, Section 5 is the conclusion of the paper and provides prospects for future work.

Related work

Previous studies have extensively investigated ML and data balancing strategies for diabetes prediction, and the findings provide useful context for interpreting the results of the present study. Huma Naz et al.[12] presented a deep learning based predictive model for detecting diabetes at an earlier stage. They employed four algorithms, namely, Deep Learning (DL), Neural Networks (NN), Naive Bayes (NB), and Decision Trees (DT). The results indicated that the deep learning strategy attained the highest accuracy compared to the other algorithms. Arief Wibowo et al.[4] have proposed A New Modified Weighted SMOTE (ANMWS) to address the class imbalance, reporting improved performance when combined with expert-driven feature selection and ML classifiers such as Support Vector Machine (SVM) & Logistic Regression (LR). Md. Maniruzzaman et al.[5] proposed a hybrid model using LR and RF to predict diabetes. LR has effectively selected the significant diabetes risk factors. In their research, they used the National Health and Nutrition Examination Survey (NHNES), consisting of 6561 instances with 657 diabetic and 5904 controls. The result shows that the hybrid of LR and RF outperformed with 94.25% accuracy.

Other studies have explored both ML and DL-based frameworks on datasets such as simulated data, local healthcare data, and the PIMA Indian Diabetes dataset (PIDD). Stacking ensemble and fused models, such as SVM, ANN integration by Md. Shamim Reza et al.[3] have shown superior accuracy. Usama Ahmed et al.[13] suggested a fused ML model to identify diabetes at an earlier stage. SVM and ANN algorithms were used in their framework. The results of these algorithms serve as the input membership function for the fuzzy model, which subsequently determines the positivity or negativity of a diabetes diagnosis. Himanshu Gupta et al.[14] suggested DL and quantum machine learning (QML) prediction models for diabetes forecasting. Deep Learning attained a superior accuracy of 95% compared to Quantum Machine Learning. Muhammad Waqas Nadeem et al.[15] presented a fusion-based ML approach using SVM and ANN with data cleaning techniques, have achieved an accuracy of 96.67% on PIMA and NHNES datasets.

Hybrid and ensemble models remain a recurring theme in diabetes research. For example, Khaled Alnowaiser[16] reported strong outcomes with a voting ensemble coupled with K-Nearest Neighbors *(*KNN*)* imputation, achieving high precision, recall, and accuracy. Hosam El-Sofany et al.[17] emphasized the automated model to control diabetes patients in Saudi Arabia. They used a ten-classifier for their experiments on PIDD datasets and the private diabetes dataset. Among all classifiers, the XGBoost with SMOTE techniques outperformed with an accuracy of 97.4%. Isfafuzzaman Tasin et al.[18] established an automated diabetes prediction method utilizing PIDD and a proprietary dataset of females from Bangladesh. SMOTE and ADASYN methodologies have been utilized to rectify the unbalanced dataset. The XGB classifier with ADASYN attained an accuracy of 81%.

Additionally, another framework is proposed that employs an explainable AI method utilizing LIME and SHAP frameworks to elucidate the model's prediction of ultimate outcomes. Gangani Dharmarathne et al.[19] developed an intuitive interface for diabetes diagnosis with machine learning. They incorporated the XGB Model with SHAP's local explanations into an interface for diabetes prediction. The PIMA Diabetes dataset was utilized for model training. The proposed explanatory interface provides

current health-related awareness for the decision taken. Pawan Whig et al.[20] presented a novel approach for diabetes classification and prediction using the Pycaret Python library. Pycaret showed that the various classifiers have different accuracies. After hyper-tuning, the authors found that Gadient Boosting achieved an accuracy of 90%.

The related work revealed that ML, DL, and hybrid ensemble methods, achieves high efficacy in diabetes classification and prediction. A number of studies reported that the class-imbalance problem is still a major issue in medical datasets, and resampling techniques, such as SMOTE, ADASYN, and other weighted techniques, have the capacity to enhance sensitivity towards minority classes.[21] Studies also emphasized the importance of proper feature selection, imputation of missing values, and robust preprocessing steps to enhance overall model performance. The boosting-based approaches, mainly XGBoost, achieved reliably good performance over the various data sets.[22] Yet another trend we see on the rise is the usage of explainable AI tools such as LIME and SHAP, which allow for interpreting model decisions, causing predictions to be less opaque, therefore contributing to more transparency in clinical pull-through. While ensemble and boosting methods have been well studied in previous works, not much attention has so far been paid to ETC using different resampling techniques. This deficiency defines the need for additional investigation on ETC-based hybrid strategies that the current work attempts to address.

MATERIALS AND METHODS

This paper presents a thorough investigation of classification and prediction of diabetes, aiming to identify better techniques for addressing imbalanced data. Biased data results in an increased error rate within the predictive model. Employing a range of statistical data preprocessing techniques enhances and improves the data quality to identify patterns and relationships within the dataset. To balance the dataset, various balancing techniques, namely down-sampling, over-sampling, and hybrid sampling techniques, are used. The subsequent subsection outlines the proposed methodological framework.

Diabetes Dataset

The Kaggle repository is the source of the Diabetes Prediction Dataset (DPD).[23] As seen in Table 1, the DPD dataset has 100000 instances with eight input variables and one output variable. Of the 100000 instances, 8500 fall into the class "True," meaning that the person has diabetes, and 91500 fall into the class "False," meaning that the person does not have diabetes.

Data Preprocessing

Preprocessing the data is a vital step to ensure data quality. Data preprocessing includes data collection, cleaning, standardization, feature selection, balancing, and representation.[24] In this stage, the 3854 duplicate rows are eliminated, and the records with undefined gender are discarded. 34% of records have no information about smoking history, which is likely a big missing value, so this column was discarded to maintain data integrity. Outliers are also eliminated to ensure clean and meaningful data. Data quality is significant because it affects the results of predictions. As a result, the dataset is now properly balanced. Further, the dataset is split into training and testing in a specific ratio, i.e., 80 and 20, respectively, so that sampling can be done effectively for better results.[25] Sampling makes a strong contribution to machine learning models because it involves choosing a representative subset of data to reliably extract features and parameters from a large dataset.[26] To maintain consistency, researchers have used sampling techniques such as up-sampling, down-sampling, and a hybrid sampling technique on the dataset. These sampling techniques perform uneven permutations and combinations of representative sets of information from the collected data. A few key points of the sampling techniques are mentioned as:

Up-sampling Technique

Up-sampling technique is the procedure of augmenting the number of samples in the minority class to match those of the majority class. Equilibrate class distribution and enhance the model's capacity to effectively learn patterns from all classes. This technique is employed to duplicate or generate new samples for the minority class.[27]

Down-sampling Technique

The down-sampling method reduces the number of samples in the most class to match the

minority class. It is a common approach in cases where one class largely dominates over the other in the case of an imbalanced dataset. It selects a random subset of samples from the majority class until it becomes the same size as the minority class. Particularly, down-sampling improves models' performance of classifiers that are sensitive to class imbalance by reducing the number of majority class samples and promoting a more balanced distribution.[28]

Hybrid Sampling Technique

The hybrid sampling technique is the combination of oversampling and undersampling techniques to solve the problem of class imbalance. The hybrid sampling approach also contributes to preventing overfitting. So, this approach results in more stable and accurate models.[29]

SMOTE-ENN

SMOTE-ENN has an integrated technique to over-sample the minority class. It's an approach to class imbalance in ML, especially when performing 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.[30] The ENN component, on the other hand, 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. By applying ENN after SMOTE, the technique refines the data space to enhance the decision boundaries of classifiers.[31] SMOTE-ENN has been shown to outperform SMOTE alone 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.[32] 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.

Extra Trees Classifier

Extra Trees Classifier (ETC) is an ensemble learner that builds on RF by adding even more randomness to the tree-building process. A random split for feature selection leads to lower variance and faster computation than standard DT or RF. RF builds a tree using bootstrap samples instead of the entire data; therefore, Extra Trees reduces bias and speeds up computation. The ETC is a powerful alternative to RF, which increases feature selection randomness and reduces the risk of overfitting.[33] As a result, it is perfect for dealing with complex decision boundaries, noisy data, and high-dimensional data. It is being used quite widely, mainly in classification problems such as financial forecasting, fraud detection, and medical diagnosis. The ETC employs Gini impurity by default is represented by Eq. (i), and entropy as an alternative for classification, while Mean Square Error and Mean Absolute Error are utilized for regression is represented by Eq. (ii).

$$Gini\ Impurity = \sum_{l=1}^{0} f_l(1 - f_l) \quad \text{...(i)}$$

$$Entopy = \sum_{l=1}^{0} - f_l\, log\ (f_l) \quad \text{...(ii)}$$

Proposed Model for Diabetes Prediction with the Lowest Error Rate

The proposed model, SEETC (SMOTE-ENN with Extra Trees Classifier), employs a structured ML pipeline for the early prediction of diabetes, as shown in Figure 1. Initially, the DPD diabetes data is taken from the Kaggle repository. The dataset has been pre-processed to eliminate duplicate rows, undefined gender information, and irrelevant features, such as smoking history, to ensure the quality of the dataset. In addition, outlier detection techniques are applied to remove outliers whose extreme values might interfere with training the models. To handle the class imbalance, the pre-processed data is further refined using the SMOTE-ENN technique, which leads to a reduction in noisy samples. Thereafter, the cleaned and balanced

dataset is split into a training and testing set in the ratio of 80:20 for better results. Subsequently, the proposed model SEETC is trained using the balanced set. Thereafter, the trained SEETC model is tested on the remaining 20% data to evaluate its predictive ability and generalization using standard performance metrics. This approach aims to provide a reliable, early diagnostic tool for diabetes prediction to extend the application range of preventive health care.

Performance Evaluation

Measuring the performance and efficiency of models is critical in guaranteeing their trustworthiness and usefulness. Various metrics help evaluate the ability of the model to correctly classify a subject as diabetic or non-diabetic while making as few Type 1 and Type 2 errors as possible. To assess model performance, we used Accuracy, $Precision_{Neg}$, $Precision_{Pos}$, $Recall_{Neg}$, $Recall_{Pos}$, $F1\_Score_{Neg}$, $F1\_Score_{Pos}$, Type 1 Error Rate & Type 2 Error Rate.

Accuracy

Accuracy is the number of correctly classified instances to the total instances. It is helpful when the dataset is balanced, but it can be misleading in class imbalance cases. Accuracy is represented by Eq. (iii).

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad ...(iii)$$

Where TruePositives (TP) refers to the count of instances correctly predicted as positive, while TrueNegatives (TN) denotes the number of cases accurately classified as negative. And FalsePositives (FP) identify the cases that were incorrectly classified as positive. Similarly, FalseNegatives (FN) represents the count of instances mistakenly categorized as negative.

$Precision_{Neg}$

$Precision_{Neg}$ quantifies the proportion of cases that were anticipated to be negative and turned out to be negative. It is represented by Eq. (iv).

$$Precision_{Neg} = TN / (TN + FN) \quad ...(iv)$$

$Precision_{Pos}$

$Precision_{Pos}$ indicates the number of cases that were genuinely positive, as opposed to the number that were predicted to be positive. It is represented by Eq. (v).

$$Precision_{Pos} = TP / (TP + FP) \quad ...(v)$$

$Recall_{Neg}$

$Recall_{Neg}$ indicates the number of genuine negatives that were accurately anticipated as negative. It is represented by Eq. (vi).

$$Recall_{Neg} = TN / (TN + FP) \quad ...(vi)$$

$Recall_{Pos}$

$Recall_{Pos}$ indicates the number of genuine positives that were accurately anticipated as positive. It is represented by Eq. (vii).

$$Recall_{Pos} = TP / (TP + FN) \quad ...(vii)$$

$F1\text{-}Score_{Neg}$

It is the harmonic mean of $Precision_{Neg}$ and $Recall_{Neg}$. It measures the ability of model to predict the negative class (non-diabetic cases) by balancing false positives and false negatives. It is represented by Eq. (viii).

$$F_1 - Score_{Neg} = 2 \times [(Precision_{Neg} \times Recall_{Neg}) / (Precision_{Neg} + Recall_{Neg})] \quad ...(viii)$$

Where $Precision_{Neg}$ is how many predicted negatives were actually negative, $Recall_{Neg}$ is how many actual negatives were correctly predicted.

$F1\text{-}Score_{Pos}$

$F1\text{-}Score_{Pos}$ is the harmonic mean of $Precision_{Pos}$ and $Recall_{Pos}$. It measures the model's accuracy in predicting the positive class (diabetic cases), which balances false positives and false negatives. It is represented by Eq. (ix).

$$F1_{Pos} = 2 \times [(Precision_{Pos} \times Recall_{Pos}) / (Precision_{Pos} + Recall_{Pos})] \quad ...(ix)$$

Where $Precision_{Pos}$ is how many predicted positives were actually positive, $Recall_{Pos}$ is how many actual positives were correctly predicted.

Type 1 Error Rate

Type 1 Error rate (T1E) is the false

positive rate. It is the proportion of instances that are truly negative that are incorrectly classified as positive. A lower T1E indicates that the less false positive predictions made by the classifier, which is crucial when the cost of a false positive is very high. The error percentage rate describes how frequently a non-diabetic gets mistakenly classified into a diabetic category. It is represented by Eq. (x).

$$T1E = FP / (FP + TN) \quad ...(x)$$

Type 2 Error Rate (T2E)

Type 2 Error rate (T2E) is the false negative rate. It is the proportion of positives that are incorrectly classified as negative. This means that having lower T2E error means that the classifier is more sensitive to identifying true positives, which is desirable in cases where missing a positive instance is more costly. Indicates how frequently a diabetic patient is misclassified as being non-diabetic. This is more important in the case of medical diagnosis to prevent undiagnosed diabetic patients than to minimize false positives. It is represented by Eq. (xi).

$$T2E = FN / (FN + TP) \quad ...(xi)$$

Table 1. Information about the Features of the Diabetes Dataset

S. No.	Feature	Explanation of Feature	Data-Type
1	gender (G)	Gender (Male/Female)	Object
2	age	Age in years	Float
3	Hypertension (Ht)	An individual has high blood pressure. (1 = Yes, 0 = No)	Integer
4	heart_disease (HD)	An Individual has any heart-related condition. (1 = Yes, 0 = No)	Integer
5	smoking_history (SH)	Describes past or current smoking habits.	Object
6	bmi	BMI is a measure of body fat based on height and weight.	Float
7	Hba1c_Level (HB)	Average blood glucose levels over the past 2 or 3 months.	Float
8	blood_glucose_level (BGL)	Shows the current level of glucose in the blood.	Integer
9	diabetes	Class variable (0: No diabetes, 1: Diabetes)	Integer

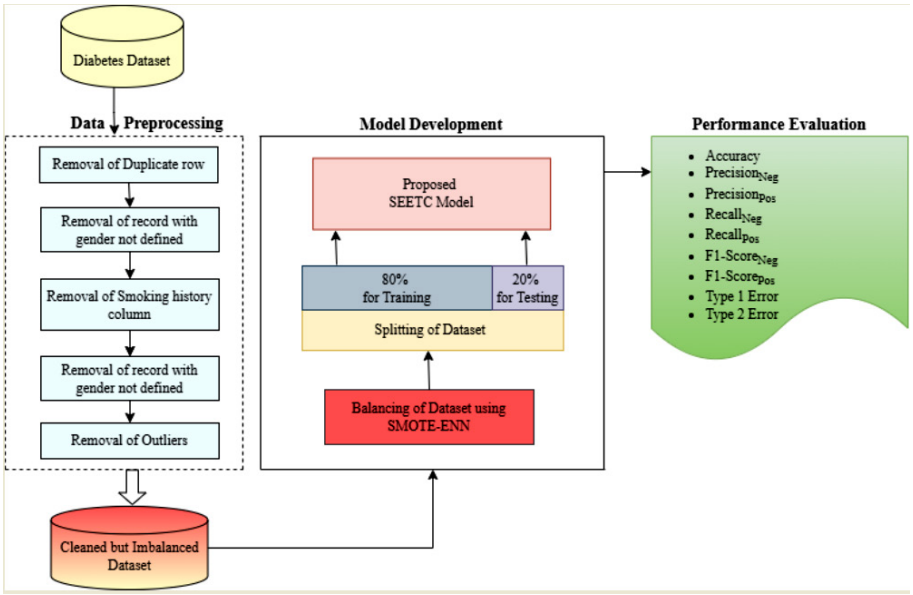

Fig. 1. SEETC Proposed Model

Selecting the right performance metrics depends on the dataset and the problem at hand. In diabetes classification, recall (sensitivity), precision (specificity), and F1-score are more important than accuracy due to class imbalance. Reducing Type 1 and Type 2 errors ensures better real-world applicability, leading to improved healthcare decision-making.

RESULTS

In this section, researchers empirically evaluate several data re-sampling techniques with the Extra Trees Classifier on the DPD dataset for diabetes prediction. We have focused primarily on two key performance metrics, Type 1 Error rate (false negative rate) and Type 2 Error rate (false positive rate). And how they influence minimizing misclassifications, which is crucial in medical diagnosis. We have measured the $Precision_{Neg}$, $Precision_{Pos}$, $Recall_{Neg}$, $Recall_{Pos}$, $F1\_Score_{Neg}$, $F1\_Score_{Pos}$, Accuracy, Type 1 Error Rate & Type 2 Error Rate for various data balancing methods to determine which model performs best. Table 2 lists a comprehensive evaluation of different resampling techniques with ETC using the DPD dataset. Among all the resampling techniques tested, ETC performed best with SMOTE-ENN and Random over-sampler. These re-sampling techniques are capable of providing the highest precise classification with a very low error rate. ETC with SMOTE-ENN exhibits the highest accuracy (0.995), F1-score (0.995) for both negative and positive classes, which exhibits an optimal balance between precision and recall, the lowest error rate (0.3) for type 1 and (0.7) for type

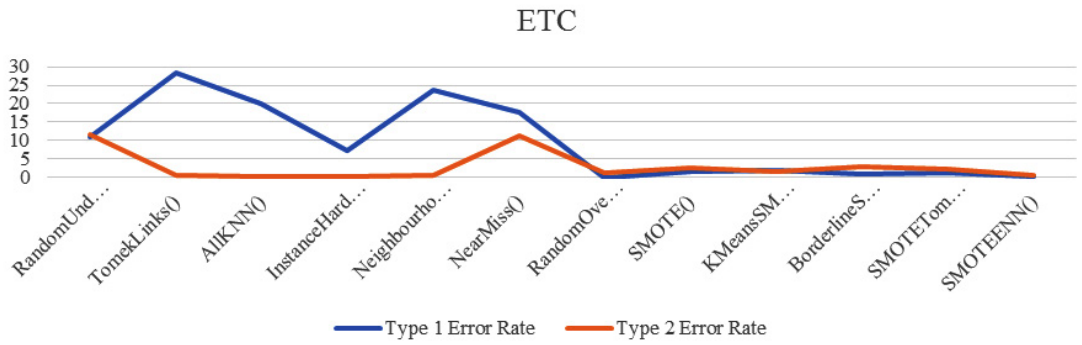

Fig. 2. Type 1 & Type 2 Error Rate using Sampling Techniques with ETC

Table 2. Impact of Re-sampling Techniques on ETC Classifier for Imbalanced Data.

S. No.	Resampling Technique	P_{Neg}	P_{Pos}	R_{Neg}	R_{Pos}	$F1_{Neg}$	$F1_{Pos}$	Acc	Type 1 Error	Type 2 Error
1	Random Under Sampler	0.893	0.884	0.883	0.893	0.888	0.889	0.888	10.9	11.7
2	Tomek Links	0.976	0.899	0.993	0.717	0.984	0.798	0.971	28.5	0.6
3	AIIKNN	0.983	0.955	0.997	0.804	0.990	0.873	0.981	20.1	0.3
4	Instance Hardness Threshold	0.992	0.972	0.997	0.928	0.995	0.949	0.990	7.4	0.3
5	Neighbourhood Cleaning Rule	0.979	0.914	0.994	0.764	0.986	0.832	0.975	23.7	0.7
6	Near Miss	0.835	0.879	0.888	0.822	0.860	0.850	0.855	17.8	11.4
7	Random Over Sampler	1	0.986	0.986	1	0.993	0.993	0.993	0.0	1.3
8	SMOTE	0.982	0.975	0.974	0.983	0.978	0.979	0.978	1.7	2.5
9	KMeans SMOTE	0.980	0.985	0.985	0.980	0.982	0.982	0.982	2.0	1.5
10	Borderline SMOTE	0.989	0.970	0.969	0.989	0.979	0.980	0.979	1	3
11	SMOTETomek	0.985	0.976	0.975	0.986	0.980	0.981	0.981	1.4	2.3
12	SMOTE-ENN	0.997	0.993	0.993	0.997	0.995	0.995	0.995	0.3	0.7

2, which indicates that the combination classified the most outcomes for diabetic patients. On the other hand, Random over-sampler with ETC also achieves high accuracy (0.993), F1-scores (0.993) for both negative and positive classes, a low error rate (0.0) for Type 1 Error, and (1.3) for Type 2 Error, whereas the remaining ten showed relatively poor performance.

This demonstrates the efficiency of the classifier with resampling techniques to enhance the predictive capabilities. The T1E and T2E rate values further highlight the ability of the proposed model. Overall, the result underscores the benefits of re-sampling techniques in improving the robustness and accuracy of the proposed model for the DPD dataset, which have been seen in Figure 2. In Figure 3, the T1E and T2E illustrate the discrimination capabilities of each resampling technique in classifying diabetes as positive or negative. The SMOTE-ENN emerges as the standout performer with a T1E of 0.3 and a T2E of 0.7.

DISCUSSION

The experimental outcomes indicate that the proposed model significantly minimizes the false-positive and false-negative error rates. Various data resampling techniques were employed to balance the dataset. The proposed model offers significant support to healthcare professionals in the early prediction of diabetes.

Impact of Resampling Techniques

Resampling techniques significantly influence the performance of ML models, especially when dealing with imbalanced datasets. In this direction, researchers have performed twelve resampling techniques. The lowest false positive and negative rates determine the accurate classification. SMOTE-ENN and Random over-sampler produced the lowest error rate, which improved the false positive and false negative rates among the other. SMOTE-ENN is the most balanced technique, keeping both Type 1 and Type 2 errors minimal.

In addition, the over-sampling techniques, including SMOTE-ENN, SMOTE, and SMOTETomek, are more effective at balancing recall, precision, Type 1 error rate, and Type 2 error rate. If false positive (Type 1 error) must be minimized, use Random Over Sampler. KMeansSMOTE and Borderline SMOTE also produce good accuracy and F1 Score, but have high error in type 2 error. While Near Miss,

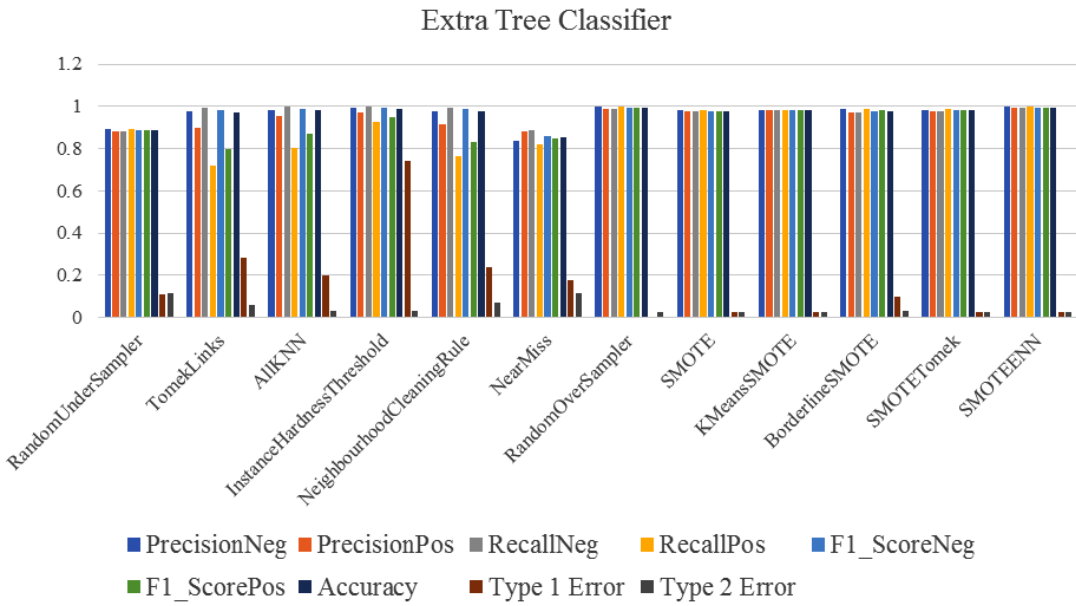

Fig. 3. Performance of different Sampling Techniques with ETC

Neighbourhood Cleaning Rule, Random Under Sampler, Instance Hardness Threshold, AIIKNN, and Tomek Links are not recommended due to significant information loss and higher errors.

Model Comparison

In this section, researchers have presented a comparative analysis to identify the most perfect model for diabetes prediction using a large dataset. Implemented a range of ML algorithms, including DT, RF, AB, XGB, and ETC, combined with the SMOTE-ENN resampling technique on the pre-processed dataset. SMOTE-ENN, which integrates both oversampling of minority classes and cleaning of noisy data using ENN, provided the most balanced and accurate results during initial experimentation, demonstrating negligible error rates. The SMOTE-ENN resampling technique performed well. The ETC consistently outperformed with SMOTE-ENN, showcasing its strength in handling high-dimensional data and feature interactions. Table 3 summarizes the performance comparison of the proposed model against the existing algorithms.

Confusion Matrix

Figure 4 shows the confusion matrix of all five ML models plotted to compare their performance. The results show that the SMOTE-ENN with Extra Trees Classifier outstandingly in data balancing and appropriate feature selection. ETC has the lowest error rate and highest accuracy among all models, which means it predicts almost accurately whether a person has diabetes or not.

Practical Implications

The proposed SEETC model provides a more reliable and practical solution to diagnose diabetes at an early stage in clinical settings. The combination of robust ensemble classifiers and hybrid data-balancing techniques makes this model suitable for clinical decision-support systems. This will facilitate the preclinical identification and monitoring of high-risk patients in healthcare. Additionally, the model's capacity to account for large patient-level datasets in a more accurate way allows deployment to lay settings. Especially in resource-limited settings where access to specialized diagnostics such as HbA1c may be limited. The model architecture is scalable and configurable to be integrated into mobile health apps and EHRs for continuous monitoring, and with a personalized care plan. The SEETC model contributes to the transition from treatment-based to a disease prevention approach and early intervention, making diabetes management more proactive and effective.

Table 3. Performance Comparison of the Proposed Model against the Existing Algorithms

S. No.	Resampling Technique	P_{Neg}	P_{Pos}	R_{Neg}	R_{Pos}	$F1_{Neg}$	$F1_{Pos}$	Acc	Type 1 Error	Type 2 Error
1	ABC	0.959	0.943	0.939	0.962	0.949	0.953	0.951	3.8	6.1
2	DT	0.984	0.983	0.982	0.985	0.983	0.984	0.984	1.5	1.8
3	XGBoost	0.980	0.991	0.990	0.981	0.985	0.986	0.985	1.9	1.0
4	RF	0.991	0.991	0.991	0.992	0.992	0.991	0.992	0.8	0.9
5	ETC	0.997	0.993	0.993	0.997	0.995	0.995	0.995	0.3	0.7

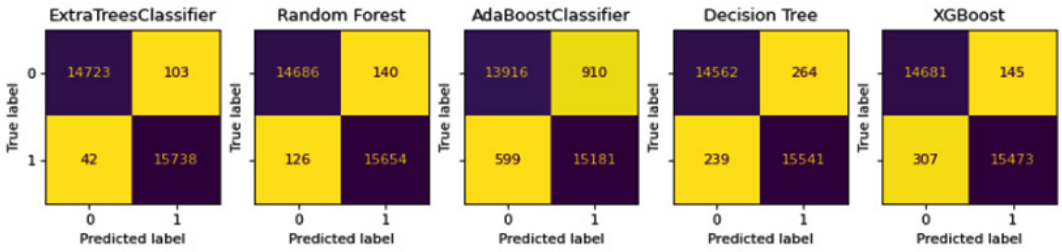

Fig. 4. Confusion Matrix

CONCLUSION

Diabetes is a chronic disease whose frequency is escalating globally. It is a passing-out vital health condition for both developed and developing countries. In this health condition, the body is unable to maintain blood sugar when the normal sugar limit is exceeded. In this paper, an early diabetes prediction system using an ensemble-based classifier and hybrid resampling techniques has been proposed. The open-source DPD has been used in this study. The SMOTE-ENN hybrid resampling technique has been applied to handle the class imbalance issues. This approach significantly improved Type I and Type II errors. Experimental outcomes showed that SMOTE-ENN, combined with the Extra Trees classifier, performed effectively in the diabetes prediction problem. The model achieved an accuracy of 99.5%, with a T1E error of less than 0.3% and a T2E error of less than 0.7%, demonstrating a very low error rate. This study also highlights the importance of choosing the right resampling technique to improve model performance when handling unbalanced data. However, a drawback of the model is that it should be tested on large real-world data from different demographic groups before assessing its potential for new populations. In future work, the model will be tested on large real-world data from different demographic groups, and will also use deep learning and explainable AI.

ACKNOWLEDGEMENT

The authors would like to acknowledge with gratitude all the support and guidance they have received from their mentors and senior PhD scholars during this study. We would also like to thank the providers of open-source datasets and tools, which were helpful in developing and evaluating the derivative model.

Funding Sources

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The author(s) do not have any conflict of interest.

Data Availability

This statement does not apply to this article.

Ethics Statement

This research did not involve human participants, animal subjects, or any material that requires ethical approval.

Informed Consent Statement

This study did not involve human participants, and therefore, informed consent was not required.

Clinical Trial Registration

This research does not involve any clinical trials.

Permission to Reproduce Material from other Sources

Not Applicable

Author Contributions

Vishal Verma performed all the experiments, conceptualized the article, planned the method, and wrote the original manuscript; Malik Shahzad Ahmed Iqbal reviewed and edited the data analysis and manuscript; Satish Kumar critiqued and revised the manuscript for clarity and scholarly content; Vandna Rani Verma and Alka Agrawal supervised the research process in general, as well as supervising and critically reviewing at all stages of the study.

REFERENCES

1. Ogle GD, Wang F, Haynes A, et al. Global type 1 diabetes prevalence, incidence, and mortality estimates 2025: Results from the International diabetes Federation Atlas, 11th Edition, and the T1D Index Version 3.0. *Diabetes Res Clin Pract*. 2025;225:112277. doi:10.1016/J.DIABRES.2025.112277
2. Abdelbaky I, Ahmed M, Taha M. Machine learning classification approaches for prediction of effective diabetes drugs. *Egyptian Informatics Journal*. 2025;31:100786. doi:10.1016/J.EIJ.2025.100786
3. Reza MS, Amin R, Yasmin R, Kulsum W, Ruhi S. Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon*. 2024;10(2). doi:10.1016/j.heliyon.2024.e24536
4. Wibowo A, Masruriyah AFN, Rahmawati S. Refining Diabetes Diagnosis Models: The Impact

of SMOTE on SVM, Logistic Regression, and Naïve Bayes. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*. 2025;7(1):197-207. doi:10.35882/JEEEMI.V7I1.596

5. 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). doi:10.1007/s13755-019-0095-z
6. Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, Sherazi HHR. Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *J Healthc Eng*. 2021;2021. doi:10.1155/2021/9930985
7. 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*. Published online July 8, 2024:96-105. doi:10.1201/9781003518587-8
8. Elgendy IA, Hosny M, Albashrawi MA, Alsenan S. Dual-stage explainable ensemble learning model for diabetes diagnosis. *Expert Syst Appl*. 2025;274:126899. doi:10.1016/J.ESWA.2025.126899
9. Matboli M, Khaled A, Ahmed MF, et al. Machine learning-based stratification of prediabetes and type 2 diabetes progression. *Diabetology & Metabolic Syndrome 2025 17:1*. 2025;17(1):227-. doi:10.1186/S13098-025-01786-6
10. Das D, Aayushman, Kumar S, Hussain MA, Reddy BR. Diabetes Prediction using Ensemble Learning Techniques. *Procedia Comput Sci*. 2025;258:3155-3164. doi:10.1016/J.PROCS.2025.04.573
11. Goswami U, Verma Y, Kar A, Kirar JS. Utility of exploratory data analysis for improved diabetes prediction. *Computational Intelligence Aided Systems for Healthcare Domain*. Published online June 14, 2023:271-294. doi:10.1201/9781003368342-12/UTILITY-EXPLORATORY-DATA-ANALYSIS-IMPROVED-DIABETES-PREDICTION-UDITA-GOSWAMI-YASHASWA-VERMA-ANWESHA-KAR-JYOTI-SINGH-KIRAR
12. Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J Diabetes Metab Disord*. 2020;19(1). doi:10.1007/s40200-020-00520-5
13. Ahmed U, Issa GF, Khan MA, et al. Prediction of Diabetes Empowered With Fused Machine Learning. *IEEE Access*. 2022;10. doi:10.1109/ACCESS.2022.3142097
14. Gupta H, Varshney H, Sharma TK, Pachauri N, Verma OP. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex and Intelligent Systems*. 2022;8(4). doi:10.1007/s40747-021-00398-7
15. Nadeem MW, Goh HG, Ponnusamy V, Andonovic I, Khan MA, Hussain M. A Fusion-Based Machine Learning Approach for the Prediction of the Onset of Diabetes. *Healthcare 2021, Vol 9, Page 1393*. 2021;9(10):1393. doi:10.3390/HEALTHCARE9101393
16. Alnowaiser K. Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model. *IEEE Access*. 2024;12. doi:10.1109/ACCESS.2024.3359760
17. El-Sofany H, El-Seoud SA, Karam OH, Abd El-Latif YM, Taj-Eddin IATF. A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App. *International Journal of Intelligent Systems*. 2024;2024. doi:10.1155/2024/6688934
18. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. *Healthc Technol Lett*. 2022;10(1-2):1. doi:10.1049/HTL2.12039
19. Dharmarathne G, Bogahawaththa M, McAfee M, Rathnayake U, Meddage DPP. On the diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence. *Intelligent Systems with Applications*. 2024;22:200397. doi:10.1016/J.ISWA.2024.200397
20. Whig P, Gupta K, Jiwani N, Jupalle H, Kouser S, Alam N. A novel method for diabetes classification and prediction with Pycaret. *Microsystem Technologies*. 2023;29(10):1479-1487. doi:10.1007/S00542-023-05473-2/FIGURES/6
21. Sneha HR, Annappa B. Exploratory Analysis of Methods, Techniques, and Metrics to Handle Class Imbalance Problem. *Procedia Comput Sci*. 2024;235:863-877. doi:10.1016/J.PROCS.2024.04.082
22. Kaur R, Gupta N. Harnessing Decision Tree-guided Dynamic Oversampling for Intrusion Detection. *Engineering, Technology & Applied Science Research*. 2024;14(5):17456-17463. doi:10.48084/ETASR.8244
23. Diabetes prediction dataset. Accessed May 6, 2025. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
24. Nandan Prasad A. Data Quality and Preprocessing. *Introduction to Data Governance for Machine Learning Systems*. Published online 2024:109-223. doi:10.1007/979-8-8688-1023-7_3
25. Torkey H, Ibrahim E, Hemdan EED, El-Sayed A, Shouman MA. Diabetes classification application

with efficient missing and outliers data handling algorithms. *Complex and Intelligent Systems*. 2022;8(1):237-253. doi:10.1007/S40747-021-00349-2/FIGURES/9

26. Susan S, Kumar A. The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. *Engineering Reports*. 2021;3(4):e12298. doi:10.1002/ENG2.12298
27. Saeed H, Ahmed M. Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology 2023 10:1*. 2023;10(1):1-10. doi:10.1186/S43067-023-00074-5
28. Sharma HS, Singh A, Chandel AS, Singh P, Sapkal ProfA. Detection of Diabetic Retinopathy Using Convolutional Neural Network. *SSRN Electronic Journal*. Published online May 17, 2019. doi:10.2139/SSRN.3419210
29. Liu Y, Zhu L, Ding L, Sui H, Shang W. A hybrid sampling method for highly imbalanced and overlapped data classification with complex distribution. *Inf Sci (N Y)*. 2024;661:120117. doi:10.1016/J.INS.2024.120117
30. Singh Y, Tiwari M. A Comprehensive Machine Learning Approach for Early Detection of Diabetes on Imbalanced Data with Missing and Outlier Values. *SN Computer Science 2025 6:3*. 2025;6(3):213-. doi:10.1007/S42979-025-03751-6
31. Omer Albasheer F, Ramesh Haibatti R, Agarwal M, Yeob Nam S. A Novel IDS Based on Jaya Optimizer and Smote-ENN for Cyberattacks Detection. *IEEE Access*. 2024;12:101506-101527. doi:10.1109/ACCESS.2024.3431534
32. Verma V, Kumar S, Verma VR, Agrawal A. An Intelligent Machine Learning Pipeline for Early Diabetes Prediction: CatBoost Ensemble with SMOTTEEN and Optuna Tuning. *International Journal of Computing and Artificial Intelligence*. 2025;6(2):229-237. doi:10.33545/27076571.2025.V6.I2C.202
33. Sharaff A, Gupta H. Extra-Tree Classifier with Metaheuristics Approach for Email Classification. *Advances in Intelligent Systems and Computing*. 2019;924:189-197. doi:10.1007/978-981-13-6861-5_17