MOBILE AND FORENSICS
Vol. 8, No. 1, March 2026, pp. 13-25
ISSN 2656-6257
13

Enhancing Early Diabetes Detection Using Tree-Based Machine Learning Algorithms with SMOTEENN Balancing

Syahrani Lonang[a,1,*], Ahmad Fatoni Dwi Putra[b,2], Asno Azzawagama Firdaus[b,3], Fahmi Syuhada[b,4], Yuan Sa'adati[b,5]

[a] Department of Information Technology, Universitas Qamarul Huda Badaruddin Bagu, Central Lombok, Indonesia
[b] Department of Computer Science, Universitas Qamarul Huda Badaruddin Bagu, Central Lombok, Indonesia
[1] lonangsyahrani3@gmail.com*; [2] ahmadfatoni@uniqhba.ac.id; [3] asnofirdaus@gmail.com; [4] fahmisy@uniqhba.ac.id; [5] yuan@uniqhba.ac.id
* corresponding author

ARTICLE INFO

Article history
Received : October 3, 2025
Revised : October 6, 2025
Accepted : Desember 20, 2025

Keywords
diabetes,
detection,
machine learning,
classification,
data balancing

ABSTRACT

Diabetes continues to be a critical global health issue, demanding accurate predictive systems to enable preventive interventions. Traditional diagnostic tests lack efficiency for large-scale early screening, which has led to growing interest in artificial intelligence solutions. This research proposed an effective methodology for diabetes classification based on tree-based algorithms enhanced with SMOTEENN balancing. The study employed the Kaggle Diabetes Prediction Dataset with 100,000 instances and eight medical and demographic features. Preprocessing steps included handling missing and duplicate values, encoding categorical variables, and scaling numerical attributes with Min-Max normalization. To address severe class imbalance, SMOTEENN was adopted, producing a cleaner and more balanced dataset. Model evaluation was performed using Stratified 5-Fold cross-validation on six classifiers: Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and CatBoost. Experimental results indicated significant gains after balancing, with ensemble methods outperforming single-tree baselines. Random Forest delivered the best overall performance (98.93% accuracy, 98.96% F1-score, 99.16% recall, 99.94% AUC), followed by CatBoost and XGBoost with comparable results above 99% AUC. While Decision Tree benefited most from SMOTEENN in relative terms, it remained less competitive. Analysis of the importance of the analysis revealed HbA1c level and blood glucose level as dominant predictors, validating clinically meaningful learning. These findings suggest that integrating hybrid resampling with ensemble tree classifiers provides reliable and general predictions for diabetes risk. The approach holds promise for deployment in healthcare decision support systems.

1. Introduction

Diabetes mellitus (DM) remains one of the most critical global health challenges, with prevalence continuously increasing and causing substantial morbidity and mortality worldwide. According to the International Diabetes Federation (IDF), approximately 537 million adults were living with diabetes in 2021, and this number is projected to rise to 643 million by 2030 [1],[2]. Beyond the human toll, diabetes contributes to severe complications including cardiovascular diseases, nephropathy, retinopathy, and neuropathy imposing a tremendous burden on healthcare systems [3]. Early and accurate detection of diabetes is crucial to prevent long-term complications and improve patient quality of life [4].

Traditional diagnostic methods such as fasting plasma glucose, oral glucose tolerance tests, and HbA1c remain the clinical standard. However, these approaches are invasive, time-consuming, and may fail to identify high risk individuals at the early stages of the disease [5]. This limitation has motivated the adoption of artificial intelligence (AI) and machine learning (ML) techniques to develop predictive models that leverage patient data for scalable, non-invasive early diabetes detection [6].

Numerous studies have applied various ML algorithms to diabetes prediction. Classical approaches such as k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) have demonstrated promising results in classification tasks [7], [8]. More advanced methods such as ensemble learning [9] and feature selection combined with dimensionality reduction [10] have further improved predictive accuracy. Tree-based algorithms such as Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost) have attracted significant attention due to their robustness, interpretability, and ability to handle nonlinear relationships in medical data [11],[12].

Despite these advancements, three critical limitations remain unaddressed, many prior studies apply SMOTE-based oversampling without addressing synthetic noise introduced during interpolation [13], [14]. This limitation is particularly problematic in clinical settings, where synthetic noise can degrade model generalization and lead to false clinical decisions. Recent research demonstrates that hybrid balancing approaches (combining oversampling and cleaning) achieve more stable results than single-resampling methods [15],[16]. Most comparative evaluations focus on 1-2 classifiers [8], [12]. A comprehensive benchmark across multiple tree-based methods is needed to guide practitioner selection.

To address these gaps, this research proposes the development of an effective tree-based machine learning model for early diabetes detection, integrating data balancing techniques to enhance classification performance. Specifically, the Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors (SMOTEENN) is employed to simultaneously reduce noise and address class imbalance more effectively than oversampling or undersampling alone. The contribution of this research is: (1) providing a comparative evaluation of multiple tree-based classifiers, including Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), AdaBoost (ADB), XGBoost (XGB), and CatBoost (CTB), (2) investigating the impact of imbalance handling, and (3) offering an interpretable model that can assist healthcare practitioners in identifying individuals at high risk of diabetes. By advancing tree-based approaches, this research seeks to improve predictive accuracy, robustness, and clinical applicability in the early detection of diabetes.

2. Method

Fig. 1 shows the diabetes detection framework proposed in this paper.

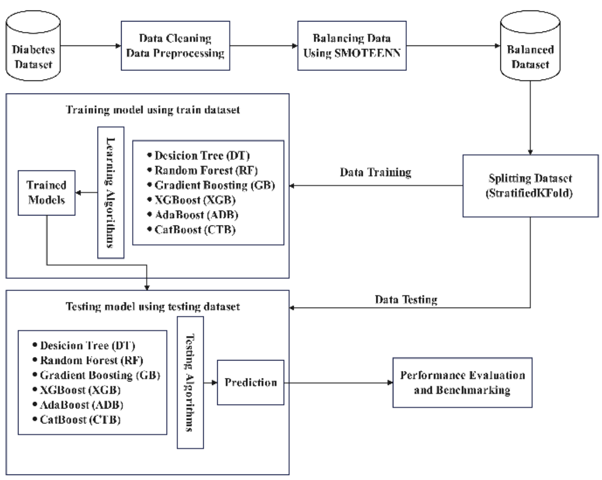

Fig. 1. The process of the proposed framework

2.1. Dataset

This research employs the Diabetes Prediction Dataset obtained from Kaggle, which provides demographic and clinical records for diabetes risk assessment. The dataset contains a total of 100,000 samples with eight features, including both numerical and categorical variables. The predictive target is diabetes status, where 0 represents non-diabetic and 1 represents diabetic cases. However, the dataset exhibits a significant imbalance, with the majority class (non-diabetic) dominating the minority class (diabetic), a common challenge in medical datasets that may reduce model sensitivity to minority outcomes [17]. Addressing this imbalance is crucial to improve the fairness and reliability of classification results. Table 1 provides a description of the features included in this dataset.

Table 1. Table Styles

Index	Feature	Data Type	Description
1	Gender	category	Female = 0; Male = 1
2	Age	Numeric	Age in years, [4,80]
3	Hypertension	Category	No hypertension =0; Having hypertension = 1
4	Heart disease	Category	No heart disease =0; Having heart disease = 1
5	Smoking history	Category	Never = 0; Other = 1; No Info = 2; Ever = 3; Current = 4; Former = 5; Not Current = 6
6	Body mass index (BMI)	Numeric	Measure of body fat based on weight and height [10,95.7]
7	HbA1c level	Numeric	Hemoglobin A1c measures a person's average blood sugar level [3.5,9]
8	Blood glucose level	Numeric	Amount of glucose in the bloodstream [80,300]
9	Diabetes	Category	No Diabetes = 0; Having Diabetes = 1

2.2. Data Preprocessing

Data preprocessing was conducted to improve data quality and model performance [18]. Missing values were inspected, and duplicate entries were removed to avoid bias. Categorical variables, including gender and smoking history, were encoded into numerical values for algorithm compatibility. Finally, all numerical features were scaled using Min-Max normalization to a range of [0,1], as expressed in Equation (1):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where x' is the normalized feature value, x is the original feature, xmin and xmax indicate the smallest and largest observed values. This method ensures uniform contribution of each feature to the learning process and avoids bias toward variables with larger ranges [19],[20].

2.3. Balancing Data Using SMOTEENN

Medical datasets are frequently imbalanced, where the majority class significantly outweighs the minority class. Such disproportionate distributions reduce classifier sensitivity and bias predictions toward the majority, thereby compromising the reliability of the model [21],[22]. The diabetes dataset employed in this research is no exception, with a stark disparity between classes. As shown in Table 2, **91,500 instances (91.5%)** belong to the non-diabetic class, while only **8,500 instances (8.5%)** represent diabetic cases. This imbalance poses a critical challenge in ensuring accurate recognition of minority outcomes.

Table 2. The distribution of diabetes in dataset

Class	Number of Records	Percentage
Non-diabetes [0]	91,500	91.5%
Diabetes [1]	8,500	8.5%
Total	100,000	100%

To address imbalance, three common strategies are employed: oversampling minority cases, undersampling majority cases, or combining both. Oversampling enriches the minority space but may induce overfitting, while undersampling reduces the dataset size and risks discarding valuable information. Prior studies have demonstrated that hybrid approaches achieve more stable and effective results compared to single resampling methods. In this research, the SMOTEENN method was adopted. SMOTE (Synthetic Minority Oversampling Technique) generates synthetic minority samples by interpolating between existing neighbors, expanding the decision boundary. ENN (Edited Nearest Neighbors) subsequently removes noisy or conflicting samples based on local neighborhood consistency, thus refining data quality. This two-step mechanism both balances and denoises the dataset [23],[24],[25]. After applying SMOTEENN, the distribution reached near parity, with 86,525 diabetic cases (51%) and 82,176 non-diabetic cases (49%). This adjustment ensures a cleaner, more balanced dataset, providing a solid foundation for reliable and generalizable model training.

Critical Data Leakage Prevention: SMOTEENN was applied exclusively within each training fold during cross-validation, never to test folds. The procedure was: 1. Dataset split into 5 stratified folds 2. For each fold: - Training set (4 folds) → Apply SMOTEENN → Train model - Test set (1-fold) → Original imbalanced distribution → Evaluate 3. Report averaged metrics across 5 folds with standard deviation. This ensures models are tested on realistic imbalanced data, simulating real-world deployment conditions where test data retain the original distribution.

2.4. Classification Model

The dataset was divided into training and testing subsets using Stratified 5-Fold cross-validation. This method ensures that each fold maintains the original class distribution, thereby reducing bias and providing a more reliable evaluation of model performance across both majority and minority classes. Such stratified sampling is widely recommended in imbalanced medical datasets to improve the stability and robustness of classification outcomes. Six tree-based algorithms were selected with the following default hyperparameters as shown Table 3:

Table 3. Hyperparameter configurations

Algorithm	Hyperparameters
Decision Tree	criterion= 'gini', max_depth=None, min_samples_split=2, random_state=42
Random Forest	n_estimators=100, criterion= 'gini', max_depth=None, min_samples_split=2, random_state=42
Gradient Boosting	n_estimators=100, learning_rate=0.1, max_depth=3, subsample=1.0, random_state=42
Algorithm	Hyperparameters
XGBoost	n_estimators=100, learning_rate=0.3, max_depth=6, subsample=1.0, random_state=42
AdaBoost	n_estimators=50, learning_rate=1.0, random_state=42
CatBoost	iterations=100, learning_rate=0.03, depth=6, random_state=42

For model development, six tree-based algorithms were selected: Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), AdaBoost (ADB), XGBoost (XGB), and CatBoost (CTB). DT: Simple yet interpretable models establishing baseline performance for medical classification tasks [26]. RF extends DT by aggregating multiple trees through bagging, which reduces variance and enhances robustness against overfitting [27]. GB and ADB represent boosting-based approaches that sequentially correct misclassified samples, often yielding higher accuracy and sensitivity in healthcare applications [28]. XGB has been widely recognized for its scalability, regularization capabilities, and superior predictive performance, making it particularly effective for handling high-dimensional medical datasets [29]. Meanwhile, CTB improves upon conventional boosting by incorporating efficient handling of categorical variables and reducing prediction bias, thus achieving competitive results even on heterogeneous clinical datasets. Prior studies in diabetes and other disease prediction have consistently shown that ensemble tree-based classifiers outperform single models in both predictive accuracy and generalization [30]. By combining these six algorithms, the research provides a comprehensive comparison of tree-based methods for early diabetes detection, highlighting relative strengths and contributions to improving diagnostic accuracy.

2.5. Performance Evaluation Measures

The evaluation of classification models in this research relies on multiple performance metrics derived from the confusion matrix, which provides a comprehensive overview of correct and incorrect predictions made by the classifier [31]. In a binary classification setting such as diabetes detection, the confusion matrix is structured around four components: true negative (TN), false negative (FN), true positive (TP), and false positive (FP). True Negative, denoted as TN in Figure. 2, shows the number of actual negative class data points predicted by the model to be actual negative. The term FN refers to the amount of data that is predicted to have a negative class but is positive. True positive, abbreviated as TP, is the amount of positive data correctly classified by the model based on its class. The amount of data in a class that is negative but is predicted to be positive by the model is referred to as a false positive (FP) [32], [22], [31].

These elements form the basis for calculating several widely used measures including accuracy, precision, recall (true positive rate), specificity (true negative rate), F1-score, miss rate, fallout, and the area under the ROC curve (AUC-ROC). Fig. 2 illustrates the confusion matrix structure.

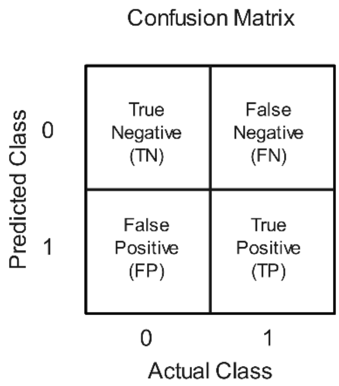

Fig. 2. Confusion matrix

Accuracy measures the overall proportion of correctly classified instances, while precision quantifies the proportion of correctly predicted positive cases among all positive predictions. Recall, or sensitivity, reflects the proportion of actual positive cases correctly identified, whereas specificity assesses the proportion of negatives correctly recognized. To balance precision and recall, the F1-score is considered as their harmonic mean [33], [34]. Metrics such as error rate, miss rate, and fallout provide additional perspectives on misclassification costs, particularly important in healthcare tasks where false negatives may lead to critical consequences [35]. Furthermore, AUC reflects how well a model predicts diabetes. Prior works on diabetes classification consistently applied these measures, confirming their suitability for assessing model robustness and clinical applicability. Based on the outcomes obtained from the confusion matrix, the performance indicators were computed using the formulas presented below.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \times 100 \quad (4)$$

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

Vol. 8, No. 1, March 2026, pp. 13-25

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Fallout = \frac{FP}{FP + TN} \tag{7}$$

$$Missrate = \frac{FN}{FN + TP} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \times 100 \tag{4}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Fallout = \frac{FP}{FP + TN} \tag{7}$$

$$Missrate = \frac{FN}{FN + TP} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \times 100 \tag{4}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Fallout = \frac{FP}{FP + TN} \tag{7}$$

$$Missrate = \frac{FN}{FN + TP} \tag{8}$$

By employing this comprehensive set of evaluation metrics, the research ensures that the models are not only optimized for accuracy but also balanced in their ability to detect minority diabetic cases while avoiding excessive false alarms.

3. Results and Discussion

This research aims to develop an effective tree-based machine learning model for early diabetes detection. To mitigate the imbalance inherent in the dataset, the SMOTEENN method was applied, ensuring a balanced representation of classes. Performance evaluation was carried out using Stratified 5-Fold cross-validation, enabling fair assessment across classes. Six tree-based algorithms were implemented in this research: DT, RF, GB, XGB, ADB and CTB.

Before evaluating model performance, this study first analyses the impact of SMOTEENN on the original class distribution of the diabetes dataset. As shown in Figure X, the raw data are highly imbalanced, with 91,500 non-diabetic instances (class 0) and only 8,500 diabetic instances (class 1), corresponding to a ratio of approximately 10.8:1. After applying SMOTEENN, the distribution becomes substantially more balanced, yielding 86,525 diabetic cases (51%) and 82,176 non-diabetic cases (49%). This near-parity distribution confirms that the hybrid resampling procedure not only increases the representation of the minority class but also removes noisy and overlapping samples from the majority class, thereby providing a cleaner and more informative dataset for subsequent model training and evaluation.

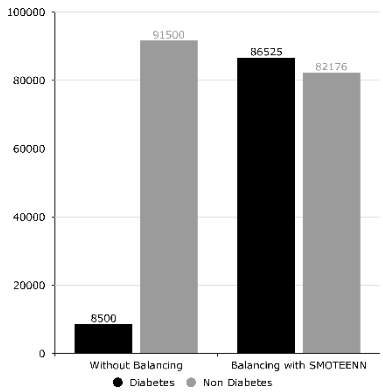

Fig. 3. Data Distribution

Table 4. Average performance for each algorithm without SMOTEENN

Algorithm	Performance Measures						
	Accuracy	Precision	Recall	F1-Score	Specificity	Fallout	Miss rate
DF	0.9519	0.7084	0.7392	0.7234	0.9717	0.0283	0.2608
RF	0.9702	0.9490	0.6860	0.7963	0.9966	0.0034	0.3140
GB	0.9720	0.9841	0.6814	0.8052	0.9990	0.0010	0.3186
XGB	0.9712	0.9552	0.6936	0.8036	0.9970	0.0030	0.3064
ADB	0.9719	1.0000	0.6691	0.8017	1.0000	0.0000	0.3309
CTB	0.9711	0.9525	0.6952	0.8037	0.9968	0.0032	0.3048

As presented in Table 4, the best overall accuracy is obtained by GB at 97.20%, closely followed by ADB (97.19%), XGB (97.12%), and CTB (97.11%). RF also performs strongly with 97.02%, while DT lags at 95.19%. This confirms the advantage of ensemble-based methods, as also reported in prior studies on disease prediction where boosting techniques consistently outperformed single-tree classifiers. For precision, ADB achieves a perfect 100.00%, indicating no false positives, while GB (98.41%) and XGB (95.52%) also attain high values. However, these models trade precision for recall, which remains comparatively low: 66.91% for ADB and 68.14% for GB. By contrast, DT shows the highest recall (73.92%) and the lowest miss rate (26.08%), reflecting its tendency to capture more true positives but at the cost of increased false alarms (lower precision, 70.84%). This inverse relationship between precision and recall is consistent with patterns highlighted in earlier works.

In terms of F1-Score, GB leads with 80.52%, slightly ahead of XGB (80.36%), CTB (80.37%), and ADB (80.17%). RF, although accurate and highly specific (99.66%), records a lower recall (68.60%), reducing its F1 (79.63%). Specificity values are uniformly high across all ensembles (≥99.66%), with ADB reaching a perfect 100.00%, but this masks the fact that minority diabetic cases are often misclassified. The results highlight that without balancing; classifiers exhibit strong precision and specificity but consistently weaker recall. This confirms the typical bias toward the majority class in imbalanced datasets, limiting the ability to correctly identify diabetic cases a challenge.

Table 5. Average performance for each algorithm with SMOTEENN

Algorithm	Performance Measures						
	Accuracy	Precision	Recall	F1-Score	Specificity	Fallout	Miss rate
DF	0.9786	0.9780	0.9803	0.9792	0.9768	0.0232	0.0197
RF	0.9893	0.9876	0.9916	0.9896	0.9869	0.0131	0.0084
GB	0.9696	0.9693	0.9714	0.9704	0.9677	0.0323	0.0286
XGB	0.9851	0.9897	0.9811	0.9854	0.9892	0.0108	0.0189
ADB	0.9462	0.9444	0.9512	0.9478	0.9410	0.0590	0.0488
CTB	0.9878	0.9928	0.9833	0.9880	0.9924	0.0076	0.0167

As reported in Table 5, the application of SMOTEENN substantially improved the performance of all classifiers compared to the imbalanced setting. The highest accuracy was achieved by Random Forest (RF) at 98.93%, followed closely by CatBoost (CTB, 98.78%) and XGBoost (XGB, 98.51%). Decision Tree (DT) also demonstrated a notable improvement, reaching 97.86%, while Gradient Boosting (GB) slightly underperformed relative to other ensembles at 96.96%. AdaBoost (ADB), however, recorded the lowest accuracy at 94.62%, showing that its conservative prediction strategy was less effective even after balancing. For precision, CTB delivered the best result (99.28%), marginally higher than XGB (98.97%) and RF (98.76%). This indicates that these ensemble methods were highly effective in minimizing false positives. Recall values also improved markedly across models, with RF attaining 99.16%, CTB 98.33%, and XGB 98.11%. Compared to the imbalanced dataset where recall values ranged between 66.91% (ADB) and 73.92% (DT), the use of SMOTEENN reduced the miss rate significantly—for instance, DT's miss rate declined from 26.08% to 1.97%, and RF's from 31.40% to 0.84%. In terms of F1-Score, RF again led with 98.96%, followed closely by CTB (98.80%) and XGB (98.54%). By contrast, ADB scored the lowest F1 (94.78%), reflecting its overall weaker balance between precision and recall despite perfect precision in the imbalanced case. Specificity values remained high across all models, with CTB attaining 99.24% and XGB 98.92%, confirming that balanced training preserved strong negative-class recognition while greatly enhancing recall score.

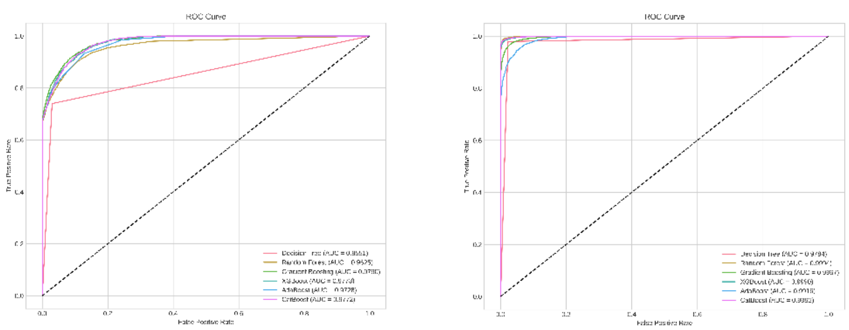

Fig. 4. The ROC curve of six algorithms with and without SMOTEENN

As presented in Fig. 4. The ROC curves reveal clear separation among models: the boosted ensembles—GB, XGB, and CTB—trace the upper-left envelope, consistent with their higher AUCs (97.90%, 97.73%, 97.72% in Table 5). RF follows with 96.25%, while DT trials markedly (85.11%), indicating limited discriminative ability when the training data are imbalanced. ADB performs well (97.28%) but remains slightly below the top trio. After balancing, every ROC curve shifts toward the top-left corner, and inter-model gaps shrink at low FPR. RF reaches the steepest ascent and the largest area (99.94% AUC), closely followed by CTB (99.92%) and XGB (99.90%). GB remains strong (99.67%). The most notable improvement is DT, whose AUC jumps from 85.11% to 97.84% (+12.73 points), reflecting far better separability of diabetic vs. non-diabetic cases once class skew is corrected. ADB also increases to 99.16%, confirming the broad benefit of hybrid resampling.

To understand which clinical and demographic attributes most strongly drive the predictions of the tree-based models, this study further examines feature importance across all classifiers, as illustrated in Fig. 5. Consistently for Decision Tree, Random Forest, Gradient Boosting, XGBoost, AdaBoost, and CatBoost, HbA1c level emerges as the most influential predictor, followed by blood glucose level, age, and BMI, whereas hypertension, heart disease, smoking history, and gender contribute comparatively less. This pattern is clinically plausible because HbA1c and blood glucose are primary biomarkers for diabetes diagnosis, while age and BMI are well-established risk factors, indicating that the models learn medically meaningful relationships rather than relying on spurious correlations.

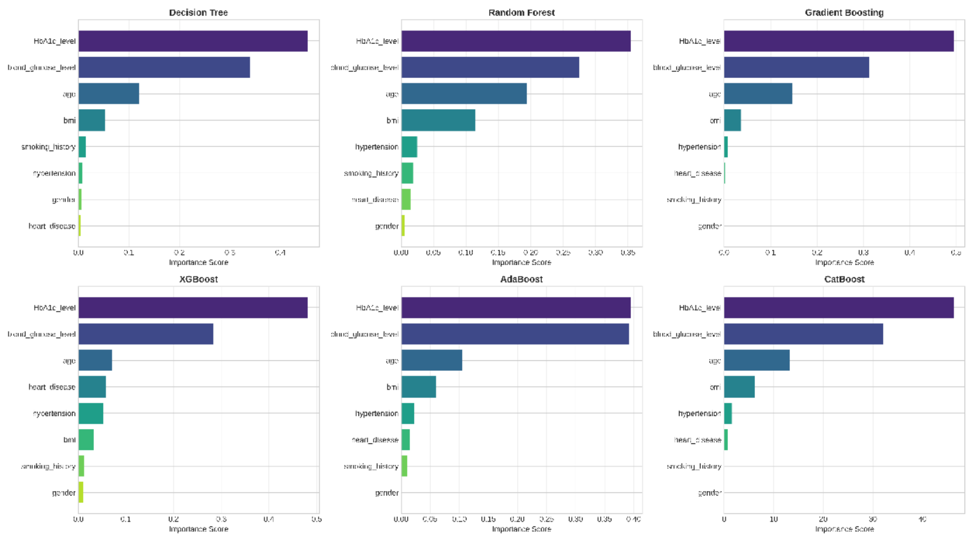

Fig. 5. Feature Importance

Considering all metrics from Tables 3–4–5 jointly, Random Forest with SMOTEENN provides the best balance: top AUC (99.94%), very high Accuracy (98.93%), Recall (99.16%), and F1-Score (98.96%), with low Fallout (1.31%) and Miss rate (0.84%). CatBoost is a close second—excelling in Precision (99.28%) and Specificity (99.24%) with AUC 99.92%—and XGBoost ranks third with consistently high Accuracy/F1 and AUC 99.90%. Thus, RF + SMOTEENN is the most effective choice for early diabetes detection in this research.

4. Conclusion

This research proposed a tree-based machine learning framework for early diabetes detection, addressing the challenge of class imbalance through the SMOTEENN technique. Six algorithms Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and CatBoost were evaluated using Stratified 5-Fold cross-validation and multiple performance indicators derived from the confusion matrix and ROC analysis.

The experimental results demonstrated that balancing with SMOTEENN produced substantial performance gains across all models, especially in recall and miss rate. Random Forest achieved the best overall performance, with an accuracy of 98.93%, F1-score of 98.96%, recall of 99.16%, and the highest AUC of 99.94%. CatBoost and XGBoost followed closely, offering strong precision, specificity, and comparable AUC values above 99.0%. Decision Tree, while showing notable improvement after balancing, remained less competitive compared to ensemble models, and AdaBoost performed relatively weaker despite gains in AUC. These findings confirm that ensemble-based classifiers combined with hybrid resampling provide a robust and reliable solution for diabetes classification.

The significance of this study lies not only in improved predictive accuracy but also in its potential clinical implications. By integrating SMOTEENN with robust classifiers, the system can support early screening and risk assessment for diabetes, thereby aiding healthcare providers in timely intervention. For future directions, the framework may be expanded with automated feature selection, hyperparameter optimization, and validation on diverse populations. Furthermore, embedding such models into digital health platforms could pave the way for scalable and real-time decision support in diabetes care.

References

[1] L. A. Al Hak, “Diabetes Prediction Using Binary Grey Wolf Optimization and Decision Tree,” *Int. J. Comput.*, vol. 21, no. 4, pp. 489–494, 2022, doi: 10.47839/ijc.21.4.2785.

[2] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, “Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction,” in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2021, pp. 141–146. doi: 10.1109/ICACCS51430.2021.9441935.

[3] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, “An Improved Artificial Neural Network Model for Effective Diabetes Prediction,” *Complexity*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/5525271.

[4] S. M. Ganie and M. B. Malik, “An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators,” *Healthc. Anal.*, vol. 2, no. January, p. 100092, 2022, doi: 10.1016/j.health.2022.100092.

[5] M. Revathi, A. B. Godbin, S. N. Bushra, and S. Anslam Sibi, “Application of ANN, SVM and KNN in the Prediction of Diabetes Mellitus,” in *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, IEEE, Apr. 2022, pp. 179–184. doi: 10.1109/ICESIC53714.2022.9783577.

[6] N. G. Ramadhan, “Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus,” *Sci. J. Informatics*, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.

[7] A. Fauzi and A. H. Yunial, “Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K – Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset,” *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 3, p. 470, Dec. 2022, doi: 10.26418/jp.v8i3.56656.

[8] P. Sugiartawan, N. W. Wardani, A. A. S. Pradhana, K. S. Batubulan, and I. N. D. Kotama, “Support Vector Machine for Accurate Classification of Diabetes Risk Levels,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 19, no. 3, pp. 317–326, Jul. 2025, doi: 10.22146/ijccs.107740.

[9] A. Dutta *et al.*, “Early Prediction of Diabetes Using an Ensemble of Machine Learning Models,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, p. 12378, Sep. 2022, doi: 10.3390/ijerph191912378.

[10] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, “Diabetes detection using deep learning techniques with oversampling and feature augmentation,” *Comput. Methods Programs Biomed.*, vol. 202, p. 105968, Apr. 2021, doi: 10.1016/j.cmpb.2021.105968.

[11] B. Aruna Devi and N. Karthik, "The Effect of Anomaly Detection and Data Balancing in Prediction of Diabetes," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 16, no. 2, pp. 184–196, 2024.

[12] Kartina Diah Kusuma Wardani and Memen Akbar, "Diabetes Risk Prediction using Feature Importance Extreme Gradient Boosting (XGBoost)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 4, pp. 824–831, Aug. 2023, doi: 10.29207/resti.v7i4.4651.

[13] A. Wibowo, A. F. N. Masruriyah, and S. Rahmawati, "Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 197–207, Jan. 2025, doi: 10.35882/jeeemi.v7i1.596.

[14] V. Sinap, "The Impact of Balancing Techniques and Feature Selection on Machine Learning Models for Diabetes Detection," *Fırat Üniversitesi Mühendislik Bilim. Derg.*, vol. 37, no. 1, pp. 303–320, Mar. 2025, doi: 10.35234/fumbd.1556260.

[15] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, "Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction," *Inf. Softw. Technol.*, vol. 139, p. 106662, 2021, doi: 10.1016/j.infsof.2021.106662.

[16] A. S. Barkah, S. R. Selamat, Z. Z. Abidin, and R. Wahyudi, "Impact of Data Balancing and Feature Selection on Machine Learning-based Network Intrusion Detection," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 241–248, 2023, doi: 10.30630/joiv.7.1.1041.

[17] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, 2022, doi: 10.1016/j.cmpb.2022.106773.

[18] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput. Math. Organ. Theory*, vol. 25, pp. 319–335, 2019, doi: 10.1007/s10588-018-9266-8.

[19] I. Riadi, A. Yudhana, and M. R. Djou, "Optimization of Population Document Services in Villages using Naive Bayes and k-NN Method," *Int. J. Comput. Digit. Syst.*, vol. 1, no. 1, pp. 127–138, 2024, doi: 10.12785/ijcds/150111.

[20] N. Alnor, A. Khleel, and K. Nehéz, "Improving the accuracy of recurrent neural networks models in predicting software bug based on undersampling methods," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 32, no. 1, pp. 478–493, 2023, doi: 10.11591/ijeecs.v32.i1.pp478-493.

[21] H. F. El-Sofany, "Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques," *IEEE Access*, vol. 12, no. July, pp. 106146–106160, 2024, doi: 10.1109/ACCESS.2024.3437181.

[22] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[23] V. Kumar, G. S. Lalotra, and R. K. Kumar, "Improving performance of classifiers for diagnosis of critical diseases to prevent COVID risk," *Comput. Electr. Eng.*, vol. 102, p. 108236, Sep. 2022, doi: 10.1016/j.compeleceng.2022.108236.

[24] C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, and S. Meng, "Advanced User Credit Risk Prediction Model Using LightGBM, XGBoost and Tabnet with SMOTEENN," *2024 IEEE 6th Int. Conf. Power, Intell. Comput. Syst. ICPICS 2024*, pp. 876–883, 2024, doi: 10.1109/ICPICS62053.2024.10796247.

[25] E. Mbunge *et al.*, "Implementation of ensemble machine learning classifiers to predict diarrhoea with SMOTEENN, SMOTE, and SMOTETomek class imbalance approaches," *2023 Conf. Inf. Commun. Technol. Soc. ICTAS 2023 - Proc.*, 2023, doi: 10.1109/ICTAS56421.2023.10082744.

[26] H. A. Abdelhafez, "Machine Learning Techniques for Diabetes Prediction: A Comparative Analysis," *J. Appl. Data Sci.*, vol. 5, no. 2, pp. 792–807, May 2024, doi: 10.47738/jads.v5i2.219.

[27] S. Saxena, D. Mohapatra, S. Padhee, and G. K. Sahoo, "Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms," *Evol. Intell.*, vol. 16, no. 2, pp. 587–603, Apr. 2023, doi: 10.1007/s12065-021-00685-9.

[28] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–24, 2023, doi: 10.1186/s12859-023-05465-z.

[29] K. D. K. Wardhani and M. Akbar, "Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost)," *J. Online Inform.*, vol. 7, no. 2, pp. 244–250, Dec. 2022, doi: 10.15575/join.v7i2.970.

[30] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[31] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Comput. Stat.*, Nov. 2022, doi: https://doi.org/10.1007/s00180-022-01301-9.

[32] D. K. Sharma, M. Chatterjee, G. Kaur, and S. Vavilala, "Deep learning applications for disease diagnosis," in *Deep Learning for Medical Applications with Unique Data*, Elsevier, 2022, pp. 31–51. doi: 10.1016/B978-0-12-824145-5.00005-8.

[33] H. Abbad Ur Rehman, C. Y. Lin, and Z. Mushtaq, "Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease," *J. Chinese Inst. Eng. Trans. Chinese Inst. Eng. A*, vol. 44, no. 1, pp. 77–87, 2021, doi: 10.1080/02533839.2020.1831967.

[34] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Diagnosing of disease using machine learning," *Mach. Learn. Internet Med. Things Healthc.*, pp. 89–111, 2021, doi: 10.1016/B978-0-12-821229-5.00003-3.

[35] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets," *J. Chinese Inst. Eng. Trans. Chinese Inst. Eng. A*, vol. 43, no. 1, pp. 80–92, 2020, doi: 10.1080/02533839.2019.1676658.

AUTHORS BIBLIOGRAPHY

SYAHRANI LONANG received his bachelor's and master's degrees in informatics from Universitas Ahmad Dahlan, Indonesia. He is currently a lecturer at the Department of Information Technology, Universitas Qamarul Huda Badaruddin Bagu. His research interests center on Artificial Intelligence, Machine Learning, Data Mining, and Data Analysis, with a particular focus on developing effective computational models to support decision-making processes. He has been actively engaged in research, and teaching aiming to integrate theoretical foundations with practical applications.

AHMAD FATONI DWI PUTRA He is a lecturer at Qamarul Huda University Badaruddin Bagu in the computer science program. He took master's degree in information technology at Mataram University, Indonesia and bachelor's degree in informatics engineering at Mataram University, Indonesia. He has a research focus on Artificial Intelligence and Internet of Things.

ASNO AZZAWAGAMA FIRDAUS He is a lecturer at Qamarul Huda University Badaruddin Bagu in the computer science program. He took master's degree in informatics at Universitas Ahmad Dahlan, Indonesia and bachelor's degree in informatics at Mataram University, Indonesia. He has research focus on Artificial Intelligence, especially on NLP, Machine Learning, Data Analysis to Data Mining.

FAHMI SYUHADA received the bachelor's degree in informatics engineering from Universitas Mataram, Indonesia, in 2017, and the master's degree in informatics engineering from Institut Teknologi Sepuluh Nopember, Indonesia, in 2020. Since 2020, he has been a Lecturer with the Department of Computer Science and Information Technology, Universitas Qamarul Huda Badaruddin, Indonesia. In 2024, he began pursuing the Ph.D. degree in computer science at Institut Teknologi Sepuluh Nopember. He has more than 30 journal articles and conference papers published. His research interests include informatics, artificial intelligence, deep learning, and image processing, and natural language processing (NLP).

YUAN SA'ADATI obtained her bachelor's degree in informatics from STMIK Lombok in 2018 and her master's degree in informatics from Universitas Islam Indonesia in 2021. Since 2021, she has been teaching at Universitas Qamarul Huda Badaruddin Bagu, where she teaches courses in Artificial Intelligence. Her research interests include the development of Decision Support Systems, machine learning algorithms, and the application of Data Science to solve real-world problems in technology and education.