Open Access ISSN: 3007-1437 (Online), ISSN: 3007-1429 (Print)

IRAQI Academic Scientific Journals

Alkadhim Journal for Computer Science (KJCS)

Journal Homepage: https://alkadhum-col.edu.iq/JKCEAS

KJCS ALKADHIM JOURNAL FOR COMPUTER SCIENCE

Explainable Artificial Intelligence Integrated Ensemble Learning Framework for Diabetes Prediction

Mohammed Abdallazez Mohammed

University of Karbala, College of Computer Science and Information Technology, Computer Science Department, Karbala, Iraq.

Article information

Article history:
Received: December, 4, 2025
Accepted:December, 13, 2025
Available online: December, 25, 2025

Keywords:
Diabetes Prediction,
Ensemble Learning,
Machine Learning,
extremely randomized Trees,
Explainable Artificial Intelligence.

Corresponding Author:
Mohammed Abdallazez Mohammed
mohammed.abdallazez@uokerbala.edu.iq

DOI:
https://doi.org/10.61710/kjcs.v3i4.141

Abstract

Accurately predicting of diabetes using clinical and demographic indicators is crucial, as early detection of this chronic metabolic disorder helps prevent serious long-term organ complications. Existing research continues to face significant challenges, including class imbalance, the scarcity of large and diverse datasets, and limited integration of explainable artificial intelligence. This research compares several ensemble learning methods including extremely randomized Trees, on a large imbalanced dataset (87,664 negative vs. 8,482 positive samples). To mitigate imbalance, we evaluate six resampling approaches including Random Over Sampling. We assess models using metrics robust to class imbalance (precision, recall, F1, AUC-ROC, and AUC-PR) and calibration measures. The Extra Trees classifier achieved the highest measured accuracy (0.994); with Random Over Sampling for balancing dataset. also, these results were compared with several previous works and number of machine learning algorithms, and the results showed superiority. Explain ability is performed at both global and local levels: permutation and SHAP for global feature importance, and (Local Interpretable Model-Agnostic Explanations) force plots for instance-level reasoning. however, we analyzed this result using sensitivity, specificity, PR-AUC and calibration, we report detailed experiments showing how resampling method, hyper parameter tuning, and stratified validation influence performance. Finally, we provide clinical-relevant insights from SHAP analyses and discuss limitations and future directions for deploying interpretable models in screening workflows.

1. Introduction

Diabetes is a long-term metabolic disease characterized by the pancreas not producing enough insulin (in type 1 diabetes, T1D) or the body's cells not responding correctly to insulin (in type 2 diabetes, T2D) [1]. Insulin is critical in the regulation of blood sugar, and the phase of metabolic homeostasis. When this regulating process breaks down and blood sugar stays high over time, a state known as hyperglycemia can cause permanent damage to the nerves or blood vessels of various organs. Diabetes has been recognized as one of the top killer diseases according to WHO [1]. likewise, 422 million people, primarily in low- and middle-income countries, live with the complaint. As a result, diabetes is responsible for approximately 1.5 million deaths per year worldwide [1]. According to the predictions made by WHO, the figure may reach approximately 642 million by 2040, which means one in a ten may suffer from diabetes as a result of bad lifestyle choices and insufficient exercise [2]. Worldwide, diabetes mellitus is a major threat to public health, and early discovery is crucial for effective care and issue prevention [3]. While there is no definitive cure, early and accurate prediction of diabetes can significantly improve patient outcomes and reduce healthcare burdens. However, predicting diabetes is challenging due to the limited availability of complex and nonlinear datasets, which often contain outliers, missing values, and unbalances [4]. Machine learning (ML) approaches have extensively used to improve early discovery and treatment [5], where early detection of diabetes, a worldwide health problem, is essential for minimizing serious complications that can occur as a result. Now, intelligent diabetic diagnostic tools with the utilization of machine learning (ML) algorithms have not only sprung up but also are given due attention in recent times. However, establishing complete correctitude in these systems is a major obstacle. Developments in ensemble ML methods indicate that it can be a promising measure to detect diabetes earlier, which will bring fast and cost-effective diagnostic modalities as compared to conventional techniques [6]. The chronic nature of diabetes demands provident strategies for early determination and intervention, indeed, in the face of improvements in operation and treatment. Prophetic modelling shows promise in fulfilling this demand, primarily through the use of ensemble learning approaches [7]. Although diabetes may be effectively prognosticated using typical prophetic modeling methods, ensemble learning presents a special opportunity to ameliorate model performance and safety. Ensemble approaches can capture intricate liaison and patterns in the data by combining prognostications from various base models, resulting in more accurate prognostications [7]. Ensemble modeling is the process of creating two or more predictive models and combining them to predict the end result. The predictions of the individual models are assembled to obtain a final prediction, which typically performs better on the test set. The objective of ensemble tree algorithms is to reduce generalization error in prediction [8]. The former is often better than isolating individual models, with two main considerations: The first is the good generalization ability and reliable predictions of ensemble models in comparison with single models. Second, in order to improve the robustness of the model and reduce variability, predictions are rendered stable and smooth over various datasets. Ensemble learning is particularly effective for data diagnosis as it addresses several limitations inherent in individual models. By combining multiple models, ensemble learning reduces prediction errors, enhances robustness against data variations, and improves generalization capabilities. It also compensates for the weaknesses of single models and increases trust in the diagnostic process. Among its key advantages are improved prediction accuracy, reduced variance, greater resistance to overfitting, and enhanced robustness. Additionally, ensemble methods offer flexibility in model selection, better handling of unbalanced datasets, and increased confidence in predictions [9]. This research aims to investigate the predictive power of ensemble learning for diabetes. A large data set will be used that contains information on clinical, demographic and lifestyle aspects will be used. Our objective is to develop and evaluate ensemble learning models capable of accurately estimating the risk of diabetes and identifying significant risk factors associated with the illness. Also, contribute to the growing body of knowledge on diabetes prediction and apply ensemble learning techniques in healthcare analytics. through the development of robust prediction models to support personalized care regimens, early identification, and, eventually, improved health outcomes for people with diabetes. This paper is an early exploration of AI in the prediction of diabetes, as well as highlighting deficiencies in existing work and directions for further research. Finally, we will abstract the key findings with regard to the objective and significance from the academic community at large. This work contributes to the following brief points:

- Develop an improved performance ensemble learning model for diabetes prediction.
- Determine the best technique to address unbalance diabetes dataset.
- Use of explainable artificial intelligence XAI to increase trustworthiness and confidence in diabetes prediction.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work and highlights existing methods relevant to this research. Section 3 presents the proposed model, including a description of the dataset, preprocessing techniques, the ML models applied, and the evaluation methodology. Section 4 discusses the experimental results and provides a detailed analysis of the findings. Section 5 provide discussion with Explainable AI , Section 6 summarize limitations and future work, Finally, Section 7 concludes the paper by summarizing the key contributions .

2. Related work

Previous work employed classical ML classifiers on structured datasets—sometimes small and homogeneous ones—leading to models suffering from overfitting with low transferability, as illustrated in Maulidiyyah et al. [10], Gupta et al. [12], and Mazhar et al. [22]. Deep learning (DL) models, including deep neural networks (DNNs), convolutional neural networks (CNNs), and autoencoders, have further improved predictive accuracy, especially in medical imaging applications, as demonstrated in Gupta et al. [12] and Olatunji [13]. but few tested the models on external or clinically validated datasets. Hybrid and ensemble methods that integrate multiple algorithms—such as CNN-LSTM, fuzzy-based systems, and ensemble frameworks using XGBoost, LightGBM, and random forests—have shown strong results in addressing class unbalance and enhancing robustness as in Ayat et al. [8], Ahmed et al. [14], Abousaber et al. [15]and Sampath et al. [16], but there have been few comparisons with the state-of-the- art DL architectures. Moreover, explainable AI techniques like SHAP, LIME, and LASSO have been incorporated to improve model interpretability as in Kaliappan et al. [17], J. Kaliappan et al. [18] and Arslan et al. [19], however, the clinical interpretability of these explanations is largely unmeasured. while emerging modalities such as ECG and voice-based prediction are broadening the application of AI in diabetes detection Jaycee M. Kaufman et al. [20] and [21]. Despite these promising developments, but they are based on small pilot datasets or unimodal input ones, without complete multimodality fusion pipelines. The existing literature is marked by several limitations (summarized in table 1.) that affect the rigor and generalizability of findings. Many studies, such as those by Maulidiyyah et al. [10] and Mazhar et al. [22], rely on small or restricted datasets, limiting statistical power and real-world applicability. External validation is frequently lacking—particularly in image-based studies like those of Olatunji et al. [13] and Ayat et al. [11]—which often evaluate models using synthetic or public datasets without clinical corroboration. Several works also exhibit narrow methodological scope, such as Sahid et al. [11], who used only traditional ML methods without exploring DL potential. Optimized or hybrid models, including those proposed by Ahmed et al. [14] and Abousaber et al. [15], were not benchmarked against state-of-the-art DL frameworks, raising concerns about performance estimation. Models trained on small, homogeneous datasets, such as Gupta et al. [12] using the PIMA Indian dataset, further risk overfitting and lack external validity. Furthermore, a substantial number of studies—e.g., Krishnan and Sheshasaayee [23]—lack interpretability components, reducing clinical usability. The absence of cross-dataset validation, as seen in works by Pratya et al. [24] and Talari et al. [25], undermines generalization across populations. Similarly, image-based studies such as Meshram et al. [26], lack comparative evaluation with clinical gold standards, and rule-based systems like Sajjadi et al. [27], though extensive in coverage, fall short in predictive strength and integration with modern AI techniques. These gaps underscore the necessity for future research to utilize diverse datasets, robust evaluation protocols, and interpretable frameworks that align with clinical practices and benchmark against contemporary AI models. Despite the growing body of literature applying machine learning and deep learning techniques to medical prediction tasks, several key knowledge gaps remain that limit the robustness, applicability, and clinical relevance of existing approaches. These gaps include:

- Most of the data sets being worked on are unbalanced, which poses a challenge.
- Many studies rely on small-scale or homogeneous datasets (e.g., PIMA Indian, LMCH), which reduces the statistical power and generalizability of the results to broader populations.
- A significant number of studies—especially image-based approaches—use only internal or synthetic datasets without clinical testing or validation in real-world healthcare environments.
- Several studies focus exclusively on traditional ML techniques or structured data, overlooking the potential of deep learning models and multimodal data integration.
- Optimized or hybrid methods are often not compared to leading DL frameworks, limiting the ability to assess their relative performance.
- Advanced models, particularly deep learning networks, are sometimes trained on limited datasets without appropriate regularization or cross-validation, increasing the risk of overfitting.
- Many ML/DL systems do not incorporate explainable AI techniques (e.g., SHAP, LIME), making them less transparent and harder to adopt in clinical practice.
- Studies often validate their models on a single dataset or population, failing to account for variability in clinical, demographic, or geographic factors.
- Some works, particularly vision-based models, are not adequately compared to established clinical diagnostic criteria or physician assessments.
- While rule-based systems use large datasets, they often neglect

- Few studies explicitly handle real-world data quality issues, such as missing values, noise, or inconsistent entries that are common in healthcare records.

In this research, have been addressed three gaps from previous works, where process the unbalanced, and using large dataset rather than on small-scale, finally incorporate explainable AI techniques.

Table (1): Comparison of limitations in related works.

Reference No.	Limitations	Details
[10],[22]	Over fitting & low transferability	Small, homogeneous datasets
[12]	Risk of over fitting	Use of PIMA Indian dataset
[12],[13]	Limited external validation	Few tested models on clinically validated datasets
[8],[14],[15],[16]	Few comparisons with state-of-the-art DL	Hybrid methods not benchmarked
[17],[18],[19]	Low measured clinical interpretability	Explainable AI explanations not clinically validated
[20],[21]	Based on small pilot/unimodal datasets	Emerging modalities lack multimodal fusion
[11],[13]	Lack of external validation	Evaluated on synthetic or public datasets
[11]	Narrow methodological scope	Only traditional ML used
[23]	Lack of interpretability	No interpretability components
[24],[25]	No cross-dataset validation	Lack of generalization across populations
[26]	Lack of comparative evaluation with clinical standards	Image-based study limitations
[27]	Weak predictive strength	Rule-based systems without modern AI integration

3. Proposed Framework

This section provides a comprehensive overview of the research procedures undertaken and the ensemble learning methods utilized during the experiments. Figure 1 illustrates the proposed framework, where involves several steps. First step the dataset undergoes preprocessing, where missing data are imputed, followed by Removing duplicate cases and formatting adjustments and Handling data set unbalance and splitting the dataset into a training set and a test set. Second step, model building, the training data was used to train the ensemble learning models (Bagging (Decision Tree), Random Forest, AdaBoost, Gradient Boosting, Histogram-based Gradient Boosting, extremely randomized trees (Ensemble), Voting Classifier (LR + RF + SVM). Third step. In the fourth step, the test phase, where model's performance is evaluated using appropriate metrics and plotting. Finally save the best model, and using XAI to interpret model behavior as a general and local explanations.

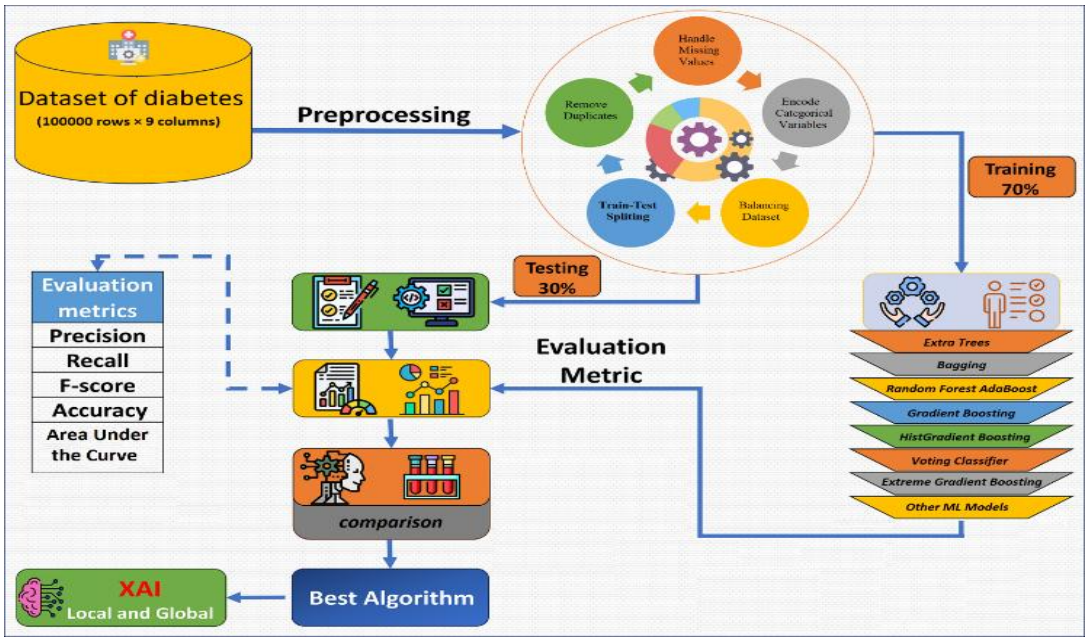

Figure (1): The flow of the proposed framework.

3.1. Dataset Overview

This dataset collects health and personal data of patients and diabetes information. Among other things, it includes age, sex, measurements of the body and blood pressure, information on heart health, smoking habits, and results from blood tests. Released for public use [28], such data aid in the construction of a tool that calculates the risk of diabetes at an individual level. This can help physicians to find out which patients might be at risk and to tailor individual treatment plans. Researchers also use this information to examine how various health and lifestyle factors are associated with developing diabetes. Shape of this dataset (100000 rows × 9 columns). The data set contains the features in Table 2.

Table (2): The Data Set Features.

Feature	Range
Gender	Male or Female
Age	0.8-80
Hypertension	0, 1
Heart-disease	0, 1
Smoking-history	Never, current, No. Info, etc.
Bmi	10 – 95.7
HbA1c-level	3.5 - 9
Blood-glucose level	80 - 300
Diabetes	Target variable, 1: has diabetes, 0: no diabetes.

3.2. Dataset Preprocessing

The data set consists of 100,000 examples with 9 columns and the preprocessing included the following steps:

- Step one: Checking Missing Value, where no missing values were found in the dataset, after conducting a complete scan to all feature's values.
- Step two: Checking Duplicated Examples, where found 3908 duplicated examples in the dataset was removed, excluding the first occurrence.
- Step three: Convert Categorical Features to Numerical, where found tow categorical features) gender, Smoking history) were, categorical to numerical distributions shown in the Table 3.
- Step four: Checking Balancing, the data set is unbalanced, where original data set shape: contains 87664 of class 0 (No Diabetes) and 8482 of class 1(Diabetes) after removing duplicated cases, this challenge was solved by applying six data balance techniques, as shown in Table 4. The training will be done on the balanced data resulting from each of the techniques in Table 3, and then the results of training will be compared to select the best technique for treating the unbalance in diabetes dataset.

Table (3): Encode Categorical Features.

Features	Numeric value	Categorical Value
Gender	0	Male
	1	Female
	2	Other
Smoking history	0	No Info
	1	Current
	2	Ever
	3	Never
	4	Former
	5	Not current

Table (4): Results of Apply Six Data-Balancing Techniques.

Techniques	Class 0	Class 1
Random Over Sampling	61397	61397
Random Under Sampling	5905	5905
SMOTE (Synthetic Minority Over-sampling Technique)	61397	61397
ADASYN (Adaptive Synthetic Sampling)	61397	61532
TomekLinks (Tomek Links Removal)	60459	5905
SMOTEENN (SMOTE + Edited Nearest Neighbours)	53408	59075

3.3. Training phase

In this section, once preprocessing is complete, the data is passed to the training process. The data were divided into 70% for training and 30% for testing, and ensemble learning (Bagging (Decision Tree), Random Forest, AdaBoost, Gradient Boosting, Histogram-based Gradient Boosting, extremely randomized Trees (Ensemble), Voting Classifier (LR + RF + SVM) and XGBoost) were applied. The Extremely randomized trees (extra tree) fall under Parallel Homogeneous Ensembles [29], where several powerful models are built using the same underlying machine learning technique. Diversity among these models is created by incorporating randomness, such as selecting different data subsets or varying input features during training. In this work, Algorithm 1 corresponds to the Extremely Randomized Trees approach, with its parameter details outlined in Table 5. this algorithm formally introduced by [30] as an additional layer of randomness in both feature and split selection for decision tree ensembles.

Algorithm 1. Extremely Randomized Trees Procedure.
Input: Training data $D = \{(x_i, y_i)\}$; number of trees M; number of random features per split K; minimum samples per leaf $n_{\min}$. For each tree $\boldsymbol{m = 1}$ to $\boldsymbol{M}$: Use the entire training dataset. At each node: (a) Randomly select Kfeatures. (b) For each selected feature, generate a random split point within the feature range. (c) Compute the information gain for classification (equation (3) based on equation (1) and (equation (2)). (d) Choose the split that yields the best score. (e) Repeat recursively until the minimum sample size $n_{\min}$is reached or no further split is possible. Combine all M trees: For classification: take the majority vote across all trees (equation (4)).

The algorithm aims to minimize an impurity measure at each node split. Common impurity measures include:

- Gini Impurity (for classification):

$$G = 1 - \sum_i (P_i)^2 \qquad (1)$$

- Entropy:

$$H = -\sum_i P_i \log_2 (P_i) \quad (2)$$

For each node, K random features are selected, and random split thresholds are generated for each. The best split is chosen based on maximum information gain, computed as:

$$\Delta I = I(parent) - \left[\frac{N_L}{N} I(left) + \frac{N_R}{N} I(right) \right] \quad (3)$$

- Final predictions:

$$\hat{y} = \text{mode}\left(T_1(x), T_2(x), \cdots T_M(x)\right) \quad (4)$$

Table 5 shows the values of the parameters passed to the algorithm with execution. which are the same as the values of the parameters passed to all other algorithms.

Table (5): Extremely Randomized Trees Parameters Values.

Parameter	Description	The value
n_estimators	Number of trees in the ensemble. More trees reduce variance but increase computation.	100
max_depth	Maximum depth of each tree. Controls model complexity and risk of overfitting.	None (default → nodes expanded until all leaves are pure or min_samples_split reached)
min_samples_split	Minimum number of samples required to split a node. Larger values prevent deep splits.	2 (default)
min_samples_leaf	Minimum number of samples required at a leaf node. Ensures leaves are not too small.	2 (default)
max_features	Number of features randomly selected at each split. Controls randomness and diversity of trees.	"sqrt" (default for classification in Extra Trees)
bootstrap	Whether bootstrap sampling is used to build each tree. If False, the full dataset is used.	False (default → Extra Trees uses the whole dataset, not bootstrapping)
random_state	Seed for random number generator to ensure reproducibility.	42
criterion	Function used to measure split quality, e.g., 'gini' or 'entropy'.	"gini" (default)
n_jobs	Number of parallel jobs used to train trees. -1 uses all available cores.	-1

Also, ensemble models like AdaBoost, Gradient boosting, histogram-based gradient boosting, XGBoost and Voting are trained on the same dataset. Additionally, A variety of have been used that have proven effective in classifying tabular medical data such as (Logistic Regression, Ridge Classifier, Passive-Aggressive, Stochastic Gradient Descent, Single-Layer Perceptron, K-Nearest Neighbors (KNN), Nearest Centroid, Support Vector Machine with Linear Kernel, Support Vector Machine with RBF Kernel, Linear Support Vector Classification, Gaussian Naive Bayes, Bernoulli Naive Bayes, Quadratic Discriminant Analysis and Linear Discriminant Analysis). Such diversity facilitates thorough benchmarking across various algorithmic families. All algorithms are trained on the same dataset as the ensemble models, and subsequently their performances are evaluated, then compare results with ensemble algorithms used.

3.4. Testing phase

Assessing how well a model works is by using appropriate evaluation criteria supporting the particular problem. The selection of these criteria depends on the properties of the dataset as well as on the kind of analysis being conducted. The fundamental evaluation metrics are presented in Table 6 for models evaluated and investigated in this research. The above measures are based on four principal ideas:

- True Positive (TP): When the model correctly predicts which patients have heart disease.
- True Negative (TN): Cases where the model correctly identifies patients without having heart disease.
- False Positive (FP): When the model wrongly predicts that a patient has heart disease when, in fact, they don't.
- False Negative (FN): It refers to the cases in which the model does not detect heart disease in patients who actually have it.

Table (6): Summarizes the Performance Metrics.

Metric	Equation
Accuracy	Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision	Precision = TP / (TP + FP)
Recall (Sensitivity)	Recall = TP / (TP + FN)
F1 Score	F1 = 2 * (precision * Recall) / (Precision + Recall)
ROC-AUC	N/A (Area under the ROC curve)

4. Results and discussion

This section presents performance metrics—including precision, accuracy, recall, F-score, ROC-AUC, and others—for several ensemble models evaluated on both unbalanced and balanced datasets. after that comparing the ensemble models with some machine learning models and the empirical research that has utilized deferent machine learning models for the prediction of diabetes disease.

4.1. Results with original dataset (unbalanced)

The performance of the designed ensemble models with the original dataset shown in learning curve with Figure 2, and Table 7. The comparison was based on precision, recall, F-score, accuracy, ROC-AUC metrics for the same size of training and testing sets for all ensemble models.

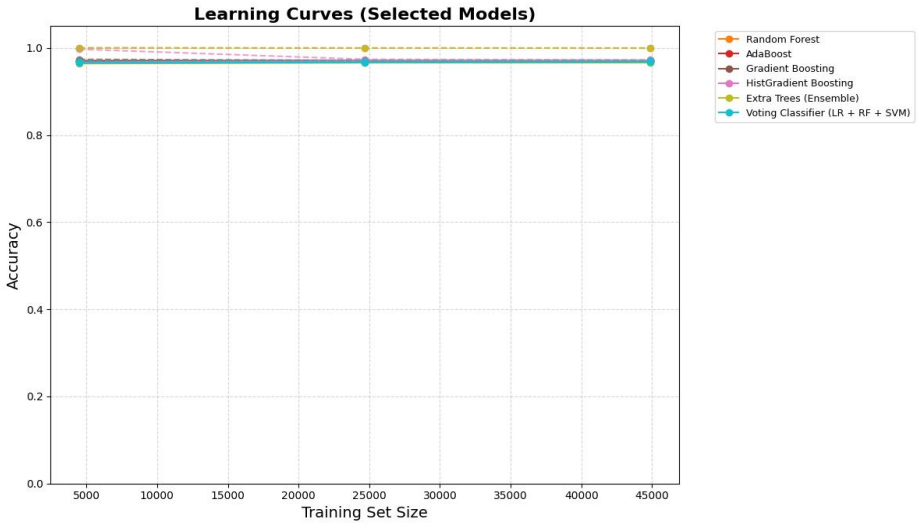

Figure (2): Learning curves with original unbalanced dataset, dashed line (validation) and solid line (training).

As we can see in Figure 1, Extremely Randomized Trees (Extra Tree) outperforms all other ensemble learning algorithms in training and validation testing.

Table (7): Results Ensemble Models with the Original Dataset (Test Set).

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Bagging (Decision Tree)	0.968	0.926	0.702	0.798	0.966
Random Forest	0.969	0.951	0.693	0.802	0.965
AdaBoost	0.971	0.989	0.688	0.812	0.976
Gradient Boosting	0.971	0.987	0.69	0.812	0.98
HistGradient Boosting	0.972	0.983	0.693	0.813	0.979
Extra Trees (Ensemble)	0.967	0.906	0.698	0.789	0.958
Voting Classifier (LR + RF + SVM)	0.969	0.962	0.682	0.798	0.972
XG Boost	0.971	0.961	0.701	0.811	0.978

The weakness of the above results is that the dataset is unbalanced and, therefore, the highest accuracy is for the highest class in the dataset. According to the accuracy scores, as shown in Table 7. HistGradient Boosting performed (0.972%). This accuracy is higher than the accuracy of all other models. The reason why the accuracy is unreliable is because the data is unbalanced. When examining the confusion matrices, we find that the error rate is very large for the class 1 as in Table 8. We note in Table 7 that the error rate in class 1 is more than 30 %, it is a big error rate. Therefore, we will work to rebalance the data and then training and testing the models to obtain a more reliable accuracy.

Table (8): Confusion Matrix of HistGradient Boosting.

Actual \ Predicted	Predicted 0 (non-diabetic)	Predicted 1(diabetic)
Actual 0 (non-diabetic)	26,236 (True Negative) 99.88%	31 (False Positive) 0.12%
Actual 1 (diabetic)	791 (False Negative) 30.69%	1,786 (True Positive) 69.31%

4.2. Results with Balanced Dataset

We rebalanced the data set using six methods as shown in Table 9. Here will display the results of training on all methods of data rebalancing, which are learning curves of Figures 3-8., and then we will analyze the training results. Figure 3 shows the results of training the ensemble learning models on balanced data using the Random Over Sampler technique, where the Extra Trees model outperformed the others, and similar results were observed in Figure 4 using Random Under Sampling. This consistent superiority was further confirmed in Figure 5 (SMOTE resampled), Figure 6 (ADASYN resampled), Figure 7 (Tomek Links resampled), and Figure 8 (SMOTEENN resampled). Extra Trees consistently outperformed the other ensemble models due to several factors: it uses fully random thresholds when splitting nodes, which reduces overfitting, creates more diverse trees, and improves generalization on unseen data; it trains faster by randomly selecting both split features and thresholds without searching for the best split, making it especially efficient for high-dimensional data; and its extreme randomness results in highly decorrelated trees, which reduces variance and leads to more stable predictions. Overall, the combination of enhanced generalization, reduced variance, and faster training made Extra Trees consistently superior across all resampling methods.

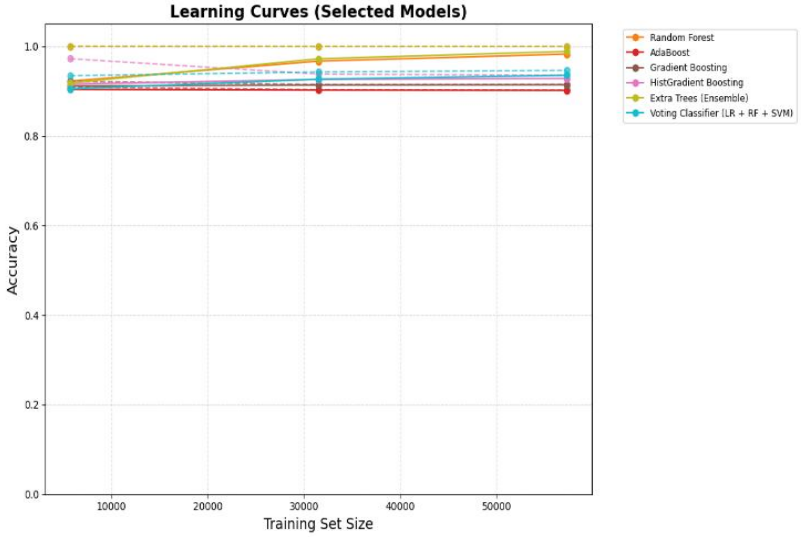

Figure (3): Ensemble training with Random Over Sampler resampled.

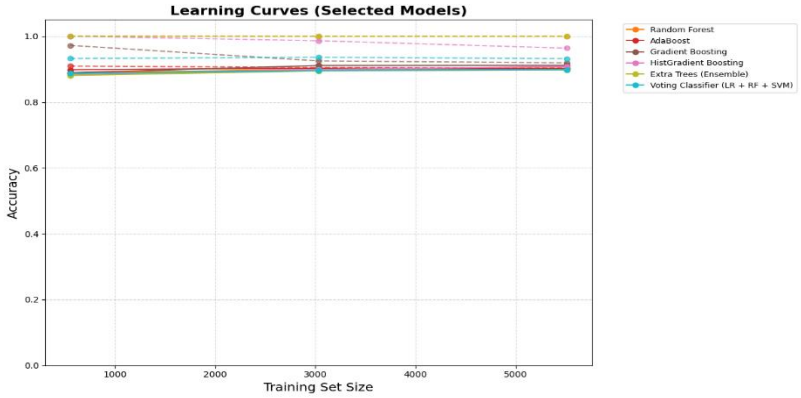

Figure (4): Ensemble training with_ Random Over Sampler resampled.

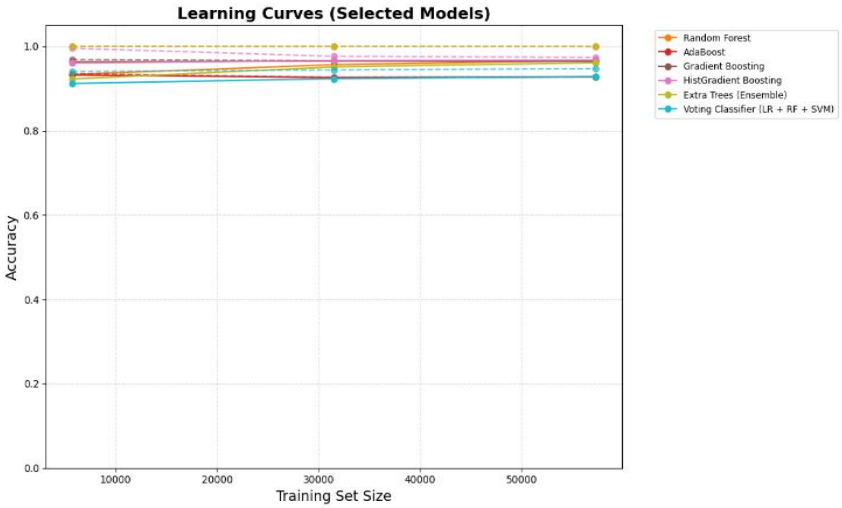

Figure (5): Ensemble training with_ SMOTE _resampled.

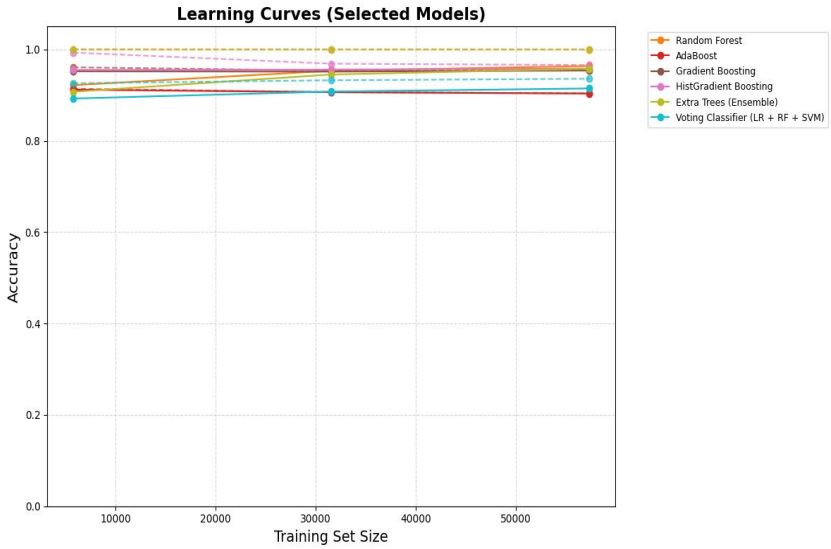

Figure (6): Ensemble training with_ ADASYN _resampled.

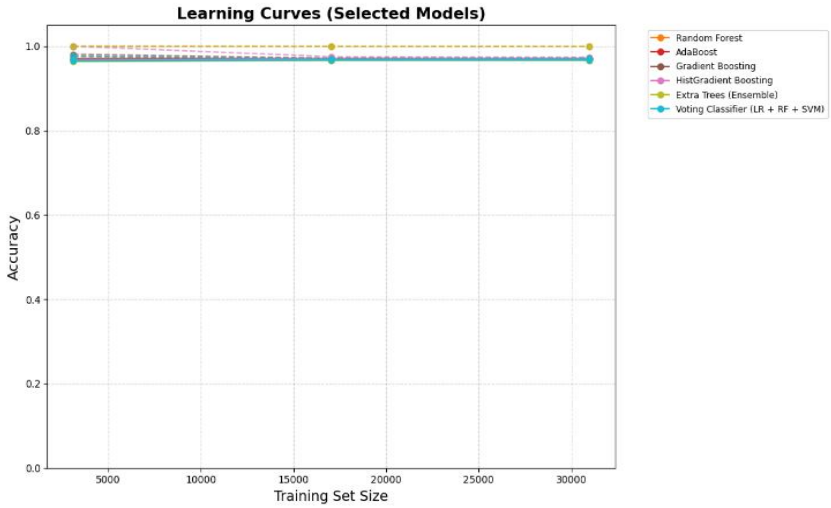

Figure (7): Ensemble training with_ Tomek Links_ resampled.

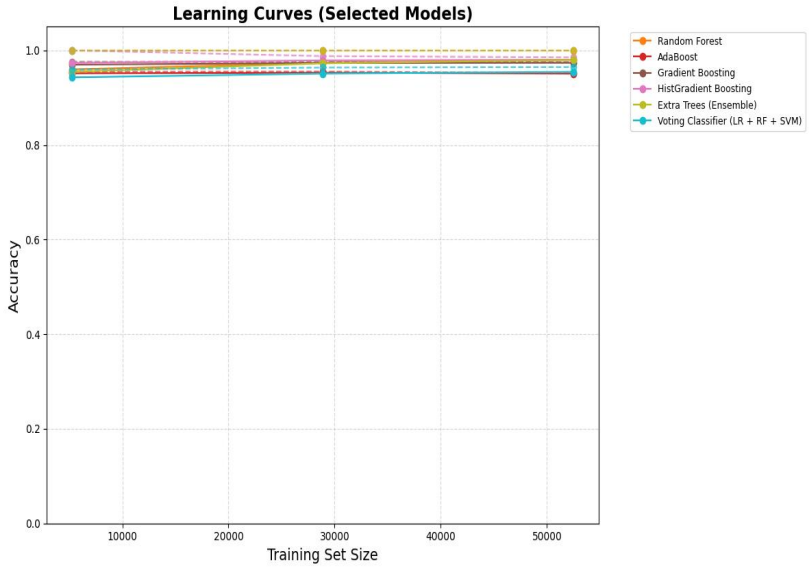

Figure (8): Ensemble training with_ SMOTEENN_ resampled.

Table (9): Results of Testing Ensemble Models with All Balanced Datasets.

Resemble techniques	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Over Sampling	Bagging (Decision Tree)	0.989	0.978	1	0.989	1
	Random Forest	0.99	0.98	1	0.99	1
	AdaBoost	0.9	0.892	0.91	0.901	0.974
	Gradient Boosting	0.915	0.902	0.93	0.916	0.98
	Hits Gradient Boosting	0.931	0.913	0.951	0.932	0.986
	Extra Trees (Ensemble)	**0.994**	**0.989**	**1**	**0.994**	**1**
	Voting Classifier (LR + RF + SVM)	0.94	0.926	0.956	0.941	0.991
	XG Boost	0.947	0.926	0.97	0.948	0.99
Random Under Sampling	Bagging (Decision Tree)	0.896	0.883	0.909	0.896	0.969
	Random Forest	0.898	0.886	0.91	0.898	0.968
	AdaBoost	0.9	0.897	0.902	0.9	0.971
	Gradient Boosting	0.908	0.895	0.923	0.909	0.975
	Hits Gradient Boosting	0.902	0.887	0.919	0.903	0.973
	Extra Trees (Ensemble)	0.897	0.886	0.909	0.897	0.963
	Voting Classifier (LR + RF + SVM)	0.895	0.883	0.908	0.895	0.967
	XG Boost	0.895	0.881	0.91	0.895	0.971
SMOTE (Synthetic Minority Over-sampling Technique)	Bagging (Decision Tree)	0.98	0.988	0.971	0.979	0.997
	Random Forest	0.972	0.974	0.97	0.972	0.997
	AdaBoost	0.928	0.924	0.932	0.928	0.986
	Gradient Boosting	0.965	0.981	0.947	0.964	0.995
	Hits Gradient Boosting	0.97	0.988	0.952	0.969	0.997
	Extra Trees (Ensemble)	0.967	0.962	0.972	0.967	0.996
	Voting Classifier (LR + RF + SVM)	0.932	0.923	0.941	0.932	0.988
	XG Boost	0.974	0.988	0.96	0.973	0.997
ADASYN (Adaptive Synthetic Sampling)	Bagging (Decision Tree)	0.979	0.986	0.972	0.979	0.997
	Random Forest	0.971	0.963	0.978	0.971	0.997
	AdaBoost	0.903	0.885	0.927	0.905	0.975
	Gradient Boosting	0.954	0.964	0.943	0.953	0.993
	Hits Gradient Boosting	0.961	0.976	0.946	0.961	0.995
	Extra Trees (Ensemble)	0.966	0.953	0.98	0.966	0.996
	Voting Classifier (LR + RF + SVM)	0.917	0.888	0.955	0.92	0.984
	XG Boost	0.966	0.977	0.954	0.966	0.996
TomekLinks (Tomek Links Removal)	Bagging (Decision Tree)	0.968	0.918	0.704	0.797	0.968
	Random Forest	0.968	0.941	0.688	0.795	0.963
	AdaBoost	0.97	0.972	0.676	0.798	0.974

	Gradient Boosting	0.97	0.97	0.685	0.803	0.978
	Hits Gradient Boosting	0.97	0.953	0.694	0.804	0.978
	Extra Trees (Ensemble)	0.967	0.915	0.693	0.788	0.96
	Voting Classifier (LR + RF + SVM)	0.967	0.938	0.669	0.781	0.969
	XG Boost	0.969	0.921	0.708	0.801	0.976
SMOTEENN (SMOTE + Edited Nearest Neighbours)	Bagging (Decision Tree)	0.984	0.987	0.983	0.985	0.999
	Random Forest	0.983	0.981	0.986	0.984	0.999
	AdaBoost	0.948	0.942	0.96	0.951	0.992
	Gradient Boosting	0.973	0.975	0.973	0.974	0.997
	Hits Gradient Boosting	0.98	0.983	0.978	0.981	0.998
	Extra Trees (Ensemble)	0.984	0.98	0.99	0.985	0.999
	Voting Classifier (LR + RF + SVM)	0.956	0.951	0.965	0.958	0.994
	XG Boost	0.981	0.985	0.979	0.982	0.999

According to the metrics in Table 8. Random Over Sampling with Extra Trees (Ensemble) give the higher than the metrics of all other models. These results after rebalancing the dataset became more realistic and reliable, as when examining the confusion matrices, we find realistic numbers for the error rates for both classes. We note that the confusion matrix of Extra Trees (Ensemble) with the resembled technique Random over Sampling as in Table 10.

Table (10): Extra Trees (Ensemble)- Confusion Metrics with Random Over Sampling.

Actual \ Predicted	**Predicted 0 (non-diabetic)**	**Predicted 1(diabetic)**
Actual 0 (non-diabetic)	18340 (True Negative) 98.91%	203 (False Positive) 1.09%
Actual 1 (diabetic)	0 (False Negative) 0.0%	18296 (True Positive) 100.00%

Note Zero error rate for class 1 (diabetic) and very small number (203) for class 0 (non-diabetic). This result is more reliable than the classification results on the data in its original, unbalanced state. The proposed model is best suited for risk scoring and early screening of diabetes, serving to identify individuals at high risk who may benefit from closer monitoring or preventive interventions. It is not intended as a standalone diagnostic tool, and clinical judgment alongside standard diagnostic tests remains essential.

4.3. Benchmarking Against other Models and Previous Studies

We note that the highest accuracy in previous studies is 99.07% in [25] while this work achieved an accuracy rate of (0.994%) with high generalization achieved through extra tree ensemble, also comparison with 14 machine learning model achieved on Random Over Sampling, balanced dataset as shown in Table 11.

Table (11): Machine Learning Models with Balanced Dataset by Random Over Sampling.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.959	0.867	0.634	0.732
Ridge Classifier	0.938	0.99	0.311	0.473
Passive Aggressive	0.921	0.564	0.523	0.543
SGD Classifier	0.954	0.802	0.644	0.714
Perceptron	0.922	0.545	0.773	0.639
K-Nearest Neighbours	0.959	0.885	0.625	0.733
Nearest Centroid	0.898	0.46	0.831	0.592
SVM (Linear)	0.959	0.918	0.593	0.72
SVM (RBF)	0.962	0.98	0.585	0.732
Linear SVC	0.959	0.897	0.609	0.725
Gaussian Naive Bayes	0.901	0.461	0.64	0.536
Bernoulli Naive Bayes	0.915	0.547	0.267	0.359
QDA	0.905	0.475	0.647	0.548
LDA	0.954	0.855	0.588	0.697
Extra Trees (Ensemble)	**0.994**	**0.989**	**1**	**0.994**

5. Discussion

In this section we will discuss the results through Explainable AI that explain and make machine learning models more interpretable and trusted have received an increasing amount of attention. These techniques are all intended to explain how models arrive at their predictions, so that people can better understand and follow the reasoning behind those predictions—i.e., explaining individual results, looking for patterns of decision-making in common, and highlighting what factors have the most influence. In this research, we use two lines of tools named LIME and SHAP to provide intuitive explanations which aid in interpreting the decisions made by models.

5.1. Global Explanations

Global explanation techniques aim to provide an overall understanding of the trained Extra Trees ensemble model, showing how the model uses input features across all predictions. Gini feature importance comes from decision trees and tree-based ensembles like Extra Trees. It is calculated based on the reduction in Gini Impurity when a feature is used to split nodes. Features that frequently reduce impurity gain higher importance scores. Table 12 show the Top Gini Feature Importances of the Diabetes dataset, where the HbA1c_level is the more importance feature and gender is the low one.

Table (12): Top Gini Feature Importance.

Feature	GiniImportance
HbA1c_level	0.321
blood_ glucose_ level	0.268
Age	0.202
Bmi	0.109
hypertension	0.036
smoking_ history	0.033
Heart_ disease	0.021
Gender	0.009

Also, Permutation importance is a model-agnostic method that measures feature importance based on performance drop. It works by shuffling each feature's values and observing how much the model's accuracy or other performance metric decreases. Features that cause larger drops are more important. Table 13 show the Top Permutation Importances of the Diabetes dataset, where the HbA1c_level is the more importance feature and heart disease is the low one.

Table (13): Top Permutation Importance.

Feature	Mean Importance	Std Importance
HbA1c_level	0.292	0.0012
blood_ glucose _level	0.254	0.0012
Age	0.131	0.0008
Bmi	0.106	0.0005
Smoking_ history	0.098	0.0008
Gender	0.071	0.0008
Hypertension	0.047	0.0008
Heart_ disease	0.027	0.0004

A surrogate decision tree approximates a black-box model's predictions and lets you see, at a glance, how the model behaves across all feature values, providing global interpretability via paths, splits, and feature importance. Where Surrogate Fidelity is 0.826 and Surrogate Max Depth is 3.0. Surrogate Fidelity, with value 0.826 means the surrogate tree correctly mimics ~82.6% of the black-box predictions. A higher fidelity means the surrogate gives a more accurate global explanation. The maximum depth of the surrogate tree. Depth limits complexity and helps maintain interpretability. the value 3.0 means the surrogate tree has three levels of splits from root to leaves, a depth of 3 gives 82.6% fidelity, which is a nice trade-off between simplicity and accuracy. Figure 9 show Surrogate tree visualization.

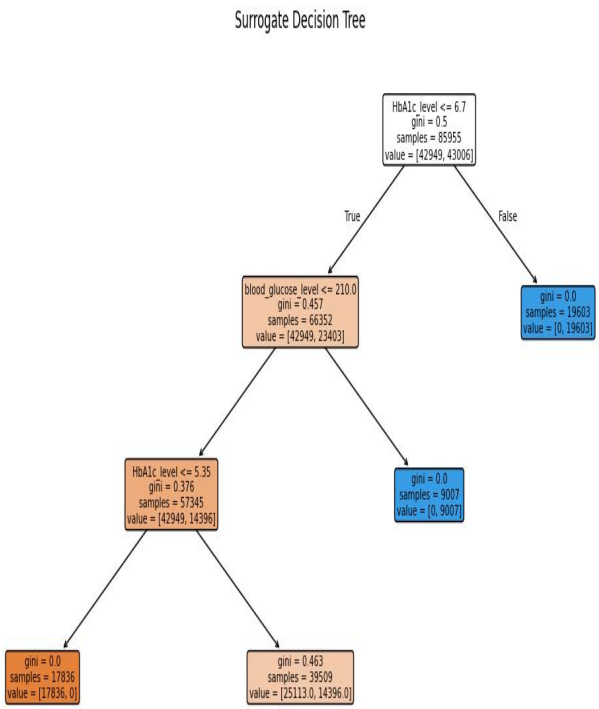

Figure (9): Surrogate tree visualization.

Figure 9. illustrates the Surrogate Tree constructed to approximate the global behavior of the trained black-box model. The surrogate tree provides an interpretable summary of how the complex model makes decisions using key clinical features, primarily HbA1c level and blood glucose level. The root node splits on HbA1c_level ≤ 6.70, indicating that this feature is the most influential variable in distinguishing between classes. When HbA1c_level is less than or equal to 6.7, the model further evaluates blood_glucose_level ≤ 210.0, forming the second major decision rule. These branches collectively represent the conditions that separate diabetic and non-diabetic predictions. The next split at HbA1c_level ≤ 5.35 refines the classification for individuals with lower HbA1c levels, further distinguishing subgroups within the non-diabetic class. Leaf nodes on the left (brown) correspond to class 0 predictions (e.g., non-diabetic), while those on the right (blue) represent class 1 predictions (e.g., diabetic). The Gini impurity values indicate the homogeneity of samples within each node, and sample counts show how many instances fall into each rule. The surrogate tree has a maximum depth of 3 and achieved a fidelity of 0.826, demonstrating that it provides a simplified yet faithful approximation of the complex model's decision process. This visualization highlights that HbA1c_level and blood_glucose_level are the key features guiding the model's global decision behavior.

5.2. Local Explanations

A local explanation in Explainable Artificial Intelligence (XAI) refers to the process of interpreting and understanding the decision-making behavior of a machine learning model for a single instance or specific prediction. local explanation focuses on identifying the most influential input features that contributed to a particular output. Local interpretability techniques used here LIME (Local Interpretable Model-Agnostic Explanations). Table 14 display some random instances feature for local explanation. This approach enables practitioners to explain why the model produced a certain prediction for an individual case, thereby enhancing model transparency, trustworthiness, and accountability in critical decision-making applications. The LIME local explanation of instances in Table 14 shown in Figures 10-14.

Table (14): Instances Feature for Local Explanation.

Instances	Gender	Age	Hypertension	Heart disease	Smoking history	Bmi	HbA1c level	Blood Glucose level
19924	0.0	3.0	0.0	0.0	0.0	17.67	4.8	80.0
41603	0.0	46.0	0.0	0.0	4.0	27.32	6.5	90.0

68027	0.0	40.0	0.0	0.0	0.0	43.21	9.0	126.0
49515	1.0	18.0	0.0	0.0	0.0	19.78	5.0	160.0
89438	1.0	51.0	1.0	0.0	1.0	31.13	6.5	160.0

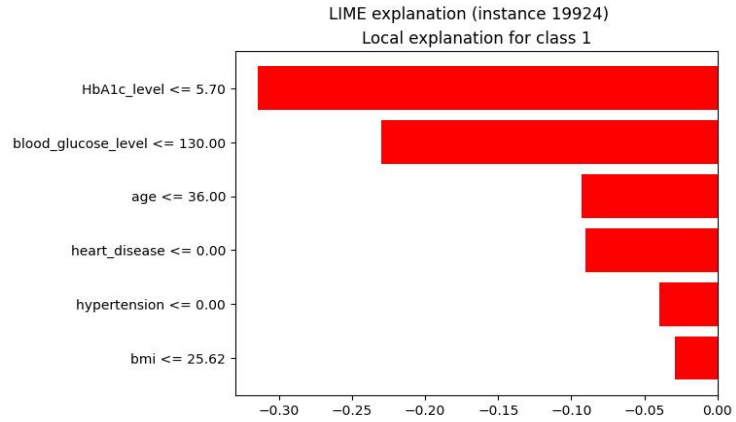

Figure (10): LIME local explanation of instance 19924.

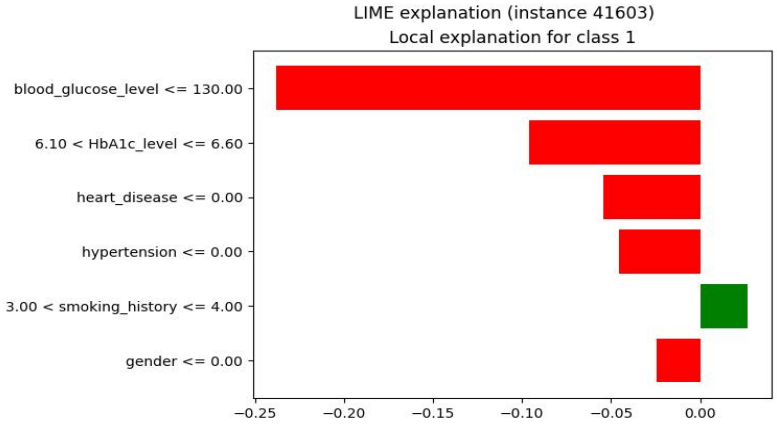

Figure (11): LIME local explanation of instance 41603.

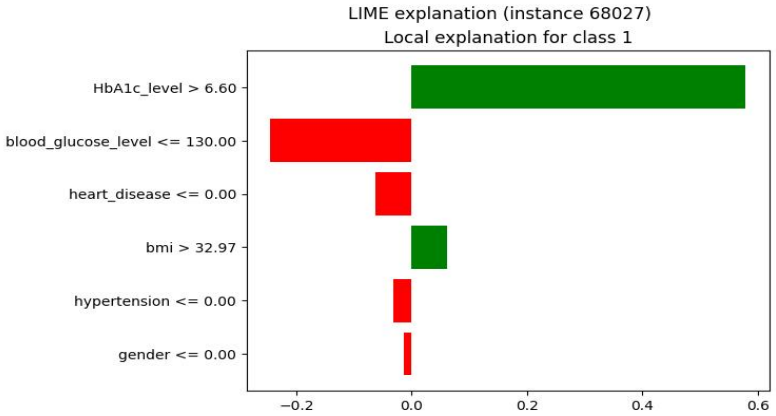

Figure (12): LIME local explanation of instance 68027.

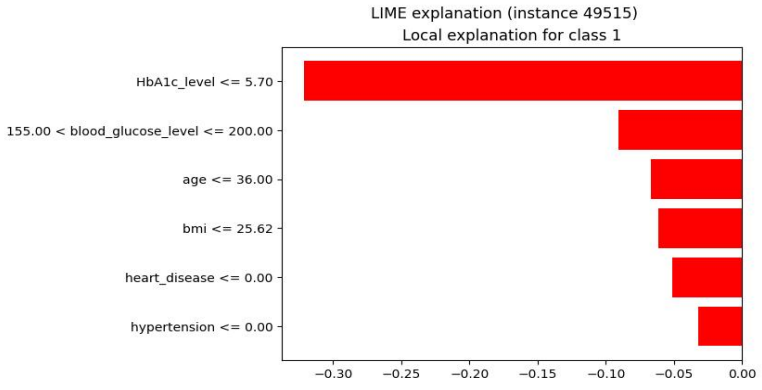

Figure (13): LIME local explanation of instance 49515.

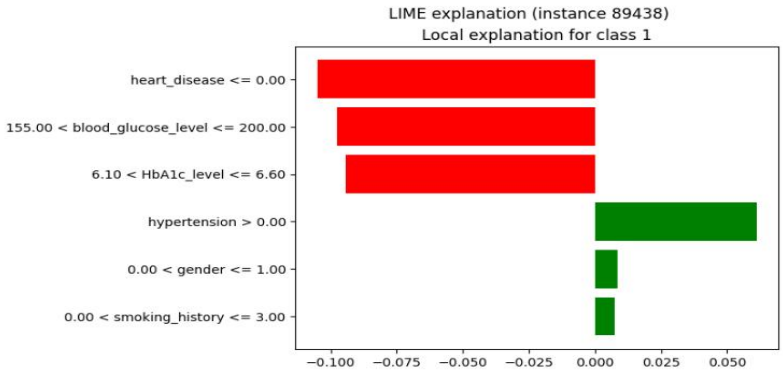

Figure (14): LIME local explanation of instance 89438.

The instance 19924, corresponding to class 1 (diabetic). in figure 10 all features exhibit negative contributions, indicating that they collectively decrease the probability of diabetes for this specific instance. The feature HbA1c_level (≤ 5.70) is the most influential factor, strongly decreasing the likelihood of being diabetic. This observation aligns with established medical guidelines, as HbA1c levels below 5.7% are generally considered normal. The blood_glucose_level (≤ 130.00) is the second most significant contributor, further reducing the model's confidence in a diabetic classification, which is consistent with healthy glucose control. The feature age (≤ 36.00) also contributes negatively, suggesting that the relatively young age of the individual reduces diabetes risk. Additionally, the absence of heart disease (heart_disease ≤ 0.00) and hypertension (hypertension ≤ 0.00) add moderate negative influence, reflecting the model's understanding that these conditions are often comorbid with diabetes. Finally, the body mass index (Bmi ≤ 25.62) provides a small negative contribution, indicating that maintaining a normal BMI further lowers the risk prediction. the local LIME explanation for instance 19924 demonstrates that the model's prediction of a low probability for diabetes is primarily driven by normal HbA1c and blood glucose levels, a young age, and the absence of cardiovascular and metabolic risk factors. This explanation confirms the model's interpretability and consistency with established medical reasoning. And with the same analytical methodology for the rest of Figures 10-14.

6. Limitations and Future Work

This study has certain limitations that need to be acknowledged. The research was conducted on a single dataset, but with a large sample size. Although the ensemble models performed well, their hyperparameters were not highly optimized. Focusing optimization on high-performing models, such as Extra Trees and XGBoost, can potentially improve robustness and accuracy. The explainability analysis, although informative, was post-hoc and highlights the need to integrate explainability into model building or build real-time explanation systems for clinicians to enable feasible deployment. Additionally, the information included static clinical measurements. Therefore, For future researches, we suggest the following:

- Using more than one dataset from multiple resources can enhance model robustness and improve generalizability, also integrating temporal information such as longitudinal biomarker trends and historical medical records may capture disease progression more effectively and yield enhanced predictive accuracy.
- Generalized of the proposed framework to heterogeneous populations, healthcare systems.
- The information included static clinical measurements. Incorporating temporal data, such as long-term biomarker trends or historical medical information, could more accurately capture disease progression and improve predictive performance.
- Optimize ensemble models hyperparameters.
- The explainability analysis, although informative, was post-hoc and highlights the need to integrate explainability into model building or build real-time explanation systems for clinicians to enable feasible deployment.
- Work large-scale validation with diverse institutions and patient cohorts to evaluate the generalizability and robustness of the proposed model.
- Investigate practical implementation aspects and Implications for Clinical Practice, together with calibration of model predictions, mitigation of fairness and bias concerns, and adaptation to population differences to ensure safe and equitable deployment in real clinical settings.

Therefore, in subsequent research, external validation will be conducted on multi-center datasets, hyperparameter tuning will be thoroughly examined, and temporal modeling strategies will be explored to capitalize on the solid foundation established in this research.

7. Conclusion

This study demonstrated that Random Over Sampling is the best method for balancing the diabetes dataset after training on a balanced dataset, where, compared six balancing techniques: Random Over Sampling, Random Under Sampling, SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), TomekLinks (Tomek Links Removal) and SMOTEENN (SMOTE + Edited Nearest Neighbors). Where eight ensemble learning techniques were used, which are (Bagging (Decision Tree), Random Forest, AdaBoost, Gradient Boosting, Histogram-based Gradient Boosting, extremely randomized Trees (Ensemble), Voting Classifier (LR + RF + SVM) and XGBoost) where evaluated by Accuracy, Precision, Recall (Sensitivity), F1 Score and ROC-AUC evaluation matrices . Extremely Randomized Trees outperformed other techniques, achieving 994% accuracy on the balanced dataset using Random Over Sampling technique because in Extra Trees, both the split thresholds and frequently the feature subsets are selected randomly. also 14 machine learning models were trained and tested for comparison with extremely randomized Trees. Explainable AI was also used to identify the strengths and weaknesses of the model and dataset.

References

[1] Olorunfemi, B.O., Ogunde, A.O., Almogren, A. et al., "Efficient diagnosis of diabetes mellitus using an improved ensemble method", *Scientific Reports*, Vol.15, No. 3235, (2025), available online: https://doi.org/10.1038/s41598-025-30733-4.

[2] Aikaterini, T., Athanasia, K., Andreas, M., "Type 2 diabetes and quality of life", *National Library of Medicine*, Vol.8, No.4, (2017), pp.120–129.

[3] Saihood, Q., Sonuç, E., "A practical framework for early detection of diabetes using ensemble machine learning models", *Turkish Journal of Electrical Engineering and Computer Sciences*, Vol.31, No.4, (2023), Article 4.

[4] Saif, D., Sarhan, A.M., Elshennaway, N.M., "Early prediction of chronic kidney disease based on ensemble of deep learning models and optimizers", *JESIT*, Vol.11, No.17, (2024).

[5] Das, D., Aayushman, Kumar, S., Hussain, M.A., Reddy, B.R., "Diabetes prediction using ensemble learning techniques", *Procedia Computer Science*, (2025).

[6] Qi, H., Song, X., Liu, S., Zhang, Y., Wong, K.K.L., "KFPredict: An ensemble learning prediction framework for diabetes based on fusion of key features", *Computer Methods and Programs in Biomedicine*, Vol.231, (2023), pp.107378.

[7] Islam, M.T. et al., "Diabetes mellitus prediction using different ensemble machine learning approaches", *2020 ICCCNT*, IEEE, (2020), pp.1–7.

[8] Ayat, Y. et al., "Novel diabetes classification approach based on CNN-LSTM: enhanced performance and accuracy", *Diagnostyka*, Vol.25, No.1, (2024).

[9] Yaman, M.A., Rattay, F., Subasi, A., "Comparison of bagging and boosting ensemble machine learning methods for face recognition", *Procedia Computer Science*, Vol.194, (2021), pp.202–209.

[10] Maulidiyyah, N.A. et al., "Comparison of decision tree and random forest methods in the classification of diabetes mellitus", *JIKO*, Vol.7, No.2, (2024), pp.79–87.

[11] Sahid, M.A., Babar, M.U.H., Uddin, M.P., "Predictive modeling of multi-class diabetes mellitus using machine learning and filtering Iraqi diabetes data dynamics", *PLOS One*, Vol.19, No.5, (2024), e0300785 .

[12] Gupta, H., Varshney, H., Sharma, T.K. et al., "Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction", *Complex Intelligent Systems*, Vol.8, (2022), pp.3073–3087.

[13] Olatunji, A., "Detection and classification of diabetic retinopathy using deep learning models", *Electronic Theses and Dissertations*, Paper 4333, (2024).

[14] Ahmed, U. et al., "Prediction of diabetes empowered with fused machine learning", *IEEE Access*, Vol.10, (2022), pp.8529–8538.

[15] Abousaber, I., Abdallah, H.F., El-Ghaish, H., "Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets", *Frontiers in Artificial Intelligence*, Vol.7, (2025), Article 1499530.

[16] Sampath, P. et al., "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique", *Scientific Reports*, Vol.14, No.1, (2024), pp.28984,.

[17] Kaliappan, J. et al., "Analyzing classification and feature selection strategies for diabetes prediction across diverse datasets", *Frontiers in Artificial Intelligence*, Vol.7, (2024), Article 1421751 .
[18] Kaliappan, J. et al., "Analyzing classification and feature selection strategies across diverse datasets", (2024).
[19] Arslan, A.K. et al., "Enhancing type 2 diabetes mellitus prediction by integrating metabolomics and tree-based boosting approaches", *Frontiers in Endocrinology*, Vol.15, (2024), Article 1444282.
[20] Jaycee, M. et al., "Acoustic analysis and prediction of type 2 diabetes mellitus using smartphone-recorded voice segments", *Mayo Clin Proc Digital Health*, Vol.1, No.4, (2023), pp.534–544.
[21] "AI could predict type 2 diabetes up to 10 years in advance", Imperial NHS, (2023).
[22] Bukhari, M.M. et al., "An improved artificial neural network model for effective diabetes prediction", *Complexity*, Vol.2021, No.1, (2021), Article 5525271.
[23] Krishnan, R.H., Sheshasaayee, A., "Optimizing diabetes classification: BOA-enhanced ML with EDA and SMOTE", (2025).
[24] Nuankaew, P., Chaising, S., Temdee, P., "Average weighted objective distance-based method for type 2 diabetes prediction", *IEEE Access*, Vol.9, (2021), pp.137015–137028.
[25] Talari, P. et al., "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2", *PLOS One*, Vol.19, No.1, (2024), e0292100.
[26] Meshram, N. et al., "Automatic detection and classification of diabetic eye disorders", *JETIR*, Vol.11, No.5, (2024).
[27] Sajjadi, S.F. et al., "Algorithms to define diabetes type using data from administrative databases: a systematic review of the evidence", *Diabetes Research and Clinical Practice*, Vol.203, (2023), pp.110859.
[28] "Kaggle diabetes prediction dataset".
[29] Kunapuli, G., "Ensemble Methods for Machine Learning", (2023), Manning Publications Co., .
[30] Geurts, P., Ernst, D., Wehenkel, L., "Extremely randomized trees", *Machine Learning*, Vol.63, No.1, (2006), pp.3–42.