IJWR

Explainable Diabetes Prediction via Hybrid Data Preprocessing and Ensemble Learning

Ghazaleh Kakavand Teimoory[a], Mohammad Reza Keyvanpour[b]*, Maryam Ghaebi[c]

[a] Data Mining Laboratory, Department of Computer Engineering Faculty of Engineering, Alzahra University Tehran, Iran; gh.kakavandteimoory@gmail.com

[b] Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran; keyvanpour@alzahra.ac.ir

[c] Data Mining Laboratory, Department of Computer Engineering Faculty of Engineering, Alzahra University Tehran, Iran; m.ghaebi@student.alzahra.ac.ir

ABSTRACT

Accurate and early prediction of diabetes is crucial for initiating prompt treatment and minimizing the risk of long-term health issues. This study introduces a comprehensive machine learning model aimed at improving diabetes prediction by leveraging two clinical datasets: the PIMA Indians Diabetes Dataset and the Early-Stage Diabetes Dataset. The pipeline tackles common challenges in medical data, such as missing values, class imbalance, and feature relevance, through a series of advanced preprocessing steps, including class-specific imputation, engineered feature construction, and SMOTETomek resampling. To identify the most informative predictors, a hybrid feature selection strategy is employed, integrating recursive elimination, Random Forest-based importance, and gradient boosting. Model training uses Random Forest and Gradient Boosting classifiers, which are fine-tuned and combined through weighted ensemble averaging to boost predictive performance. The resulting model achieves 93.33% accuracy on the PIMA dataset and 98.44% accuracy on the Early-Stage dataset, outperforming previously reported approaches. To enhance transparency and clinical applicability, both local (LIME) and global (SHAP) explainability methods are applied, highlighting clinically relevant features. Furthermore, probability calibration is performed to ensure that predicted risk scores align with true outcome frequencies, increasing trust in the model's use for clinical decision support. Overall, the proposed model offers a robust, interpretable, and clinically reliable solution for early-stage diabetes prediction.

Keywords— Diabetes Prediction; Explainable AI; Ensemble Learning; LIME; SHAP; E-Health.

1. Introduction

Diabetes is one of the fastest-rising health threats worldwide in the 21st century [1]. Diabetes occurs when energy metabolism is disrupted due to insufficient insulin production, complete lack of insulin, or the body's resistance to insulin [2]. Diabetes is generally classified into three main types, with type 2 being particularly significant due to its partially preventable nature. Type 2 diabetes results from a combination of genetic susceptibility and lifestyle factors. While certain gene mutations increase the risk, lifestyle choices largely determine when the disease manifests [3]. Type 2 diabetes primarily affects individuals aged 40 to 65 [4] and is often delayable through lifestyle changes. Typical symptoms may involve extreme thirst, increased urination, persistent tiredness, unexpected weight or muscle loss, delayed wound healing, and impaired vision [5], high blood pressure, and abnormal cholesterol levels. Over 18% of global deaths are linked to cardiovascular disease, cancer, chronic respiratory conditions, and diabetes, underscoring a significant public health concern [6]. According to the International Diabetes Federation, around 537 million people globally had diabetes in 2021 [7]. As type 2 diabetes progresses, it can lead to severe complications such as heart disease, stroke, neuropathy, kidney failure, and vision loss [8]. Statistics indicate that approximately 11% of global deaths are attributable to diabetes, highlighting the urgent need for early detection and proactive management of type 2 diabetes mellitus (T2DM) [9]. Due to the lack of noticeable symptoms, type 2 diabetes can go undetected or unmanaged for extended periods, substantially increasing the

doi http://dx.doi.org/10.22133/ijwr.2025.526792.1292

Citation G. Kakavand Teimoory, M.R. Keyvanpour, M. Ghaebi, "Explainable Diabetes Prediction via Hybrid Data Preprocessing and Ensemble Learning ", *International Journal of Web Research*, vol.8, no.4,pp.51-66, 2025, doi: http://dx.doi.org/10.22133/ijwr.2025.526792.1292.

*Corressponding Author
Article History: Received: 31 May 2025; Revised: 26 September 2025; Accepted: 29 September 2025.

financial burden on both patients and healthcare systems [10].

Machine learning and data mining methods are playing a growing role in public health initiatives and research [11]. Advancements in machine learning have addressed many limitations of traditional methods in disease diagnosis and prevention. Machine learning algorithms, in particular, can integrate a wider range of variables and data types, enabling them to generate more generalizable results than traditional statistical methods [12]. Nevertheless, several challenges persist. Deep learning methods offer substantial improvements in disease prediction by revealing hidden patterns and insights that are often difficult to detect through traditional means. However, their practical application in medical contexts remains limited owing to the opaque nature of their decision-making processes in clinical and diagnostic contexts [13]. This challenge underscores the importance of Explainable Artificial Intelligence (XAI) [14], which aims to clarify how models analyze data and make decisions. While traditional machine learning methods often allow for inherent interpretability, they may struggle to match the complexity and accuracy of more advanced models, making it difficult to fully understand their internal processes. Therefore, it is crucial to employ these predictive methods alongside interpretability techniques to enhance both their effectiveness and clinical applicability [15].

This study introduces a comprehensive and integrated machine learning model for predicting diabetes. The proposed pipeline tackles major challenges in clinical data analysis, including missing values, class imbalance, and feature relevance. Effective data preprocessing is critical for optimizing machine learning model performance [16]. It incorporates advanced data preprocessing, innovative feature engineering, and hybrid feature selection techniques to optimize predictive accuracy. To boost reliability, we apply cutting-edge ensemble learning algorithms alongside thorough model tuning. The resulting approach achieves both high performance and interpretability, making it well-suited for real-world clinical decision support.

The key contributions of this study are as follows:

- We present a machine learning pipeline for diabetes prediction that integrates class-wise, feature-specific imputation with hybrid feature selection.
- We design an ensemble modeling framework that combines Random Forest and Gradient Boosting classifiers, with performance enhanced through weighted ensemble averaging guided by cross-validated ROC-AUC scores.

This manuscript is an extended and significantly enhanced version of our earlier conference paper [17]. While the conference version primarily focused on feature selection for diabetes prediction, the present work introduces several important extensions. These include: (i) an additional clinical dataset for external validation, (ii) advanced class-specific imputation using Random Forest Regression, (iii) novel feature engineering strategies combined with hybrid feature selection, (iv) data balancing through SMOTETomek, (v) a weighted ensemble of Random Forest and Gradient Boosting classifiers, and (vi) interpretability using LIME and SHAP. Together, these additions provide a more comprehensive, robust, and explainable predictive framework compared to the initial study.

The following sections of this paper are arranged as described below:

Section 2 defines the problem, and Section 3 outlines the Related Works. Section 4 details the pipeline components: 4.1 Datasets, 4.2 Data Preprocessing, 4.3 Model Training, 4.4 Ensemble Prediction, 4.5 Explainability. Section 5 presents the Results, Section 6 discusses the findings, and Section 7 concludes the study and outlines future work.

2. Problem Definition

In this study, we formalize diabetes prediction as a supervised binary classification problem as Equation (1) denotes the collection of patient records, where each x_i is a d-dimensional feature vector representing clinical and demographic attributes and Equation (2) denotes the corresponding set of labels, where $y_i \in \{0,1\}$ indicates the absence or presence of diabetes, respectively. The primary objective is to learn a mapping Equation (3) that accurately predicts the diabetes status for previously unseen patients.

$$X = \{x_i\}_{i=1}^{N} \tag{1}$$

$$Y = \{y_i\}_{i=1}^{N} \tag{2}$$

$$f^*: \mathbb{R}^d \to \{0,1\} \tag{3}$$

3. Related Work

Extensive research has been conducted on predicting diabetes using machine learning and deep learning methods, with a focus on improving accuracy and reliability. Issues like uneven class distribution and lack of model explainability, and optimization have led to the development of innovative approaches that enhance prediction performance and enable more accurate predictions. The following section reviews studies that contribute to the advancement of diabetes prediction methods.

Reza et al. [1] introduced an enhanced non-linear kernel for Support Vector Machines to improve type 2 diabetes prediction using the PIMA dataset. The proposed kernel integrates Radial Basis Function (RBF) and RBF City Block kernels, enabling SVM to handle complex decision boundaries and improve prediction accuracy. The model outperformed traditional kernels in various evaluation metrics. However, the study relies on a synthetic over-sampling approach to address class imbalance and requires further validation across diverse populations to confirm its generalizability. Tasin et al. [7] developed an automated system for diabetes prediction leveraging machine learning and XAI techniques. They employed a variety of classifiers, including Decision Tree, SVM, Random Forest, and XGBoost, with SMOTE and ADASYN techniques for class imbalance and explainable AI methods to improve model interpretability. Agliata et al. [18] developed a model with a shallow neural network trained on a large, balanced dataset generated from multiple public sources. The model employed ensemble learning and optimization techniques to improve performance. While accurate, its simple architecture may limit its ability to capture subtle time-dependent patterns. Hama Saeed [19] investigated type 2 diabetes classification using Decision Tree, AdaBoost, Gradient Boosting, and extra trees classifiers. The study employed an up-sampling technique to address class imbalance and evaluated performance on both the PIMA and BRFSS datasets. Among the models, the extra trees classifier showed the highest accuracy, outperforming others in both datasets. Farsana and Poulose [5] introduced a hybrid convolutional neural network model integrating traditional machine learning classifiers, leveraging both spatial feature extraction and conventional learning strengths. This approach demonstrated improved prediction performance but requiring high computational resources. Tanim et al. [4]introduced DeepNetX2, a custom deep neural network integrating XAI techniques, specifically Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) using a tailored feature selection method based on Spearman correlation coefficients, achieving high accuracy and improved interpretability across multiple diabetes datasets. While the model enhances transparency and diagnostic performance, its increased architectural complexity may lead to overfitting, especially on smaller or imbalanced datasets. Its advantageous time complexity, however, helps offset this added depth and complexity, contributing to its overall effectiveness. Zhou et al. [20] proposed a generalized diabetes prediction framework that applies within and between datasets, as well as blending parts of different datasets, using ensemble and deep learning models with robust preprocessing. This approach improves robustness and generalizability across different populations. Shams et al. [21] created a hybrid model for diabetes prediction that combines Recursive Feature Elimination (RFE) to select the most important features with a GRU neural network to improve diabetes classification on the PIMA Indian dataset. This combination not only boosted accuracy but also made the model easier to understand by focusing on key features and addressing gradient problem-solving. Talukder et al. [22] implemented a machine learning approach for diabetes prediction using the PIMA dataset, emphasizing robust data preprocessing techniques such as missing value imputation, outlier removal, and SMOTE-based data balancing. Among the five models tested, XGBoost achieved the highest accuracy of 92%, showcasing the effectiveness of ensemble learning and feature management. The framework relies solely on a single dataset and lacks detailed analysis on model interpretability. Bhat et al. [23] applied six machine learning algorithms on the PIMA dataset, using optimized feature selection techniques, CFS, SFS, and Information Gain, to improve type 2 diabetes prediction. The Decision Tree classifier achieved the highest accuracy of 96.10%, but the model is limited in generalizability as it was trained on one dataset and lack of evaluation on real-time or diverse population data.

4. Proposed Method

This section details the proposed machine learning pipeline for diabetes prediction. An overview of the entire framework is illustrated in Figure 1. The workflow starts by loading the dataset, followed by an extensive preprocessing phase that involves data exploration, imputation of missing values, feature engineering, scaling, balancing, and feature selection. During model training, the data is divided into training and testing subsets, and two ensemble models, Random Forest and Gradient Boosting, are trained with hyperparameter optimization. The predictions from both classifiers are then merged using a weighted averaging method. The final classification result identifies whether the individual is diabetic or not. The LIME framework is employed to improve the transparency and interpretability of the model’s predictions.

4.1. Datasets

This study utilizes two benchmark datasets to evaluate the performance of the proposed model. The PIMA Indians Diabetes Dataset [24] is one of the most extensively used datasets for developing and evaluating diabetes prediction models. Initially developed by the National Institute of Diabetes and Digestive and Kidney Diseases, it focuses on female PIMA Indian individuals aged 21 and older, a group with a high predisposition to type 2 diabetes. The dataset includes 768 records, each containing eight key medical attributes: number of pregnancies, blood glucose level, blood pressure, skinfold thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable is binary, indicating whether or not the individual has

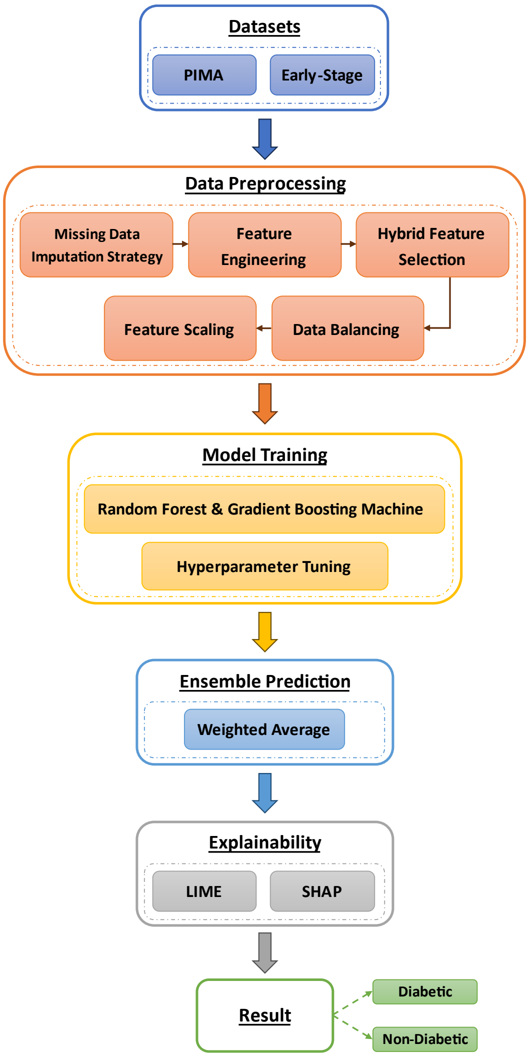


Figure. 1. Proposed method diagram

diabetes. Due to its structured format and clinical relevance, the dataset has become a preferred resource for researchers developing machine learning models for early diabetes diagnosis. Table 1 shows summary of features in the PIMA dataset, including clinical and biometric attributes used for diabetes prediction. Figure 2 shows the feature distributions in the PIMA Diabetes Dataset. Variables like insulin and skin thickness display right-skewed patterns, whereas glucose, BMI, and blood pressure appear more normally distributed. These histograms highlight data variability and inform preprocessing decisions. Figure 3 displays boxplots of features in the PIMA Diabetes Dataset, highlighting the central tendency, spread, and presence of outliers across variables.

The Early-Stage Diabetes Dataset [25] was created to facilitate the detection of diabetes during its initial stages, before the onset of severe symptoms. Collected through patient questionnaires at Sylhet Diabetes Hospital in Bangladesh, the dataset contains 520 entries and 16 features. These include both prominent and subtle symptoms, such as polyuria, polydipsia, sudden weight loss, weakness, genital thrush, and blurred vision, as well as demographic details like age and gender. The target label classifies individuals as either diabetic (positive) or non-diabetic (negative). This dataset serves as a valuable resource for developing machine learning tools aimed at early diabetes detection, particularly in settings where conventional diagnostic tests are costly or inaccessible. Table 2 presents an overview of the attributes in the Early-Stage dataset, reflecting patient symptoms and demographic characteristics relevant

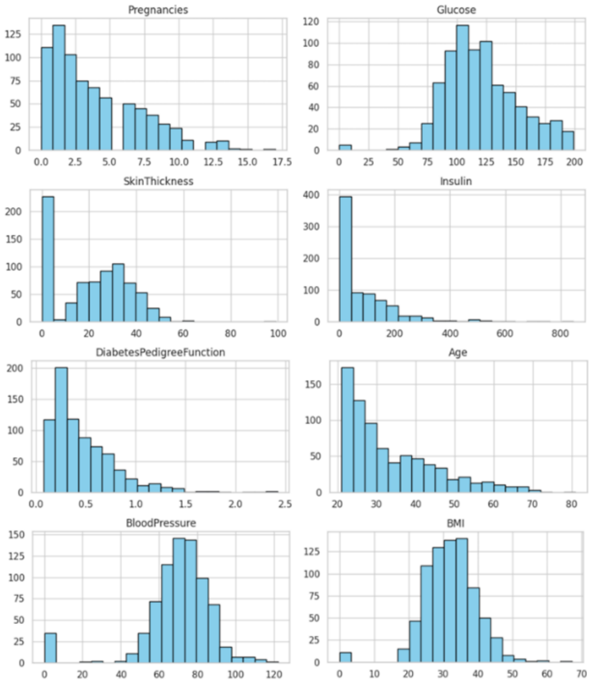


Figure. 2. Histogram of PIMA Dataset features

Table 1. Summary of features in the PIMA Dataset

Feature	***Range***
Pregnancies	0 – 17
Glucose	0 – 199
BloodPressure	0 – 122
SkinThickness	0 – 99
Insulin	0 – 846
BMI	0 – 67.1
DiabetesPedigreeFunction (DPF)	0.078 – 2.42
Age	21 – 81
Outcome	0 or 1

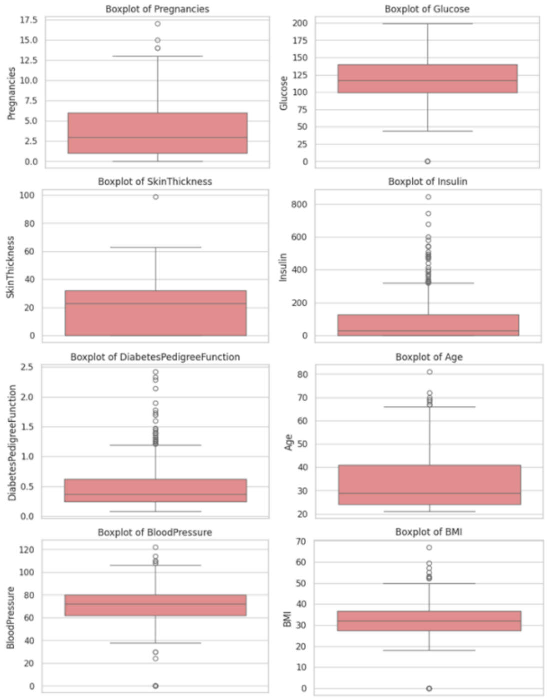


Figure. 3. Boxplot of features in PIMA Dataset

to early diabetes diagnosis. Figure 4 presents how symptoms and demographics is distributed in the Early-Stage Diabetes Dataset by class. Notable differences between diabetic and non-diabetic groups are observed in features like polyuria, sudden weight loss, and weakness, highlighting their importance in early detection.

4.2. Data Preprocessing

Preprocessing data is essential for improving the effectiveness of any machine learning model [26], especially in medical applications. In this study, we applied thorough preprocessing to support accurate diabetes prediction. Since every feature can carry valuable information, handling missing values was a priority. We then applied feature engineering, leveraging medical knowledge to create additional meaningful variables. To identify the most predictive attributes, a hybrid feature selection method was used, recognizing that each feature plays a distinct role in the outcome. Finally, we balanced the dataset to avoid model bias toward the majority class.

Missing Data Imputation Strategy

Missing data frequently occur in clinical and public health research, and machine learning–based imputation methods, particularly those using Random Forest algorithms, are increasingly being adopted [27]. On the other hand, the presence of missing values in this data set is obvious. Therefore, it was decided to replace the values instead of removing them to avoid affecting the model. Given that feature distributions can vary between diabetic and non-diabetic groups, we apply imputation separately for each class. For features with missing values, an Iterative Imputer with a Random Forest Regressor is used independently on each subset. This model-driven strategy captures intra-class variable relationships to produce more accurate and context-specific imputations. By maintaining class-specific data patterns and reducing bias, this method improves the overall quality of the dataset. The imputation strategy employed in this study enhances data quality by preserving class-specific feature distributions and capturing complex, non-linear relationships among clinical variables through Random Forest Regression. Unlike simpler methods such as mean imputation or Linear Regression, this model-based approach is more robust to outliers and noise, which are common in medical datasets. The iterative nature of the process allows missing values to be progressively refined using updated information from other features, resulting in more accurate and unbiased

Table 2. Overview of attributes in the Early-Stage Diabetes Dataset

Attribute	***Description***
Age	Age group of the patient
Gender	Biological sex
Polyuria	Excessive urination
Polydipsia	Excessive thirst
sudden weight loss	Recent unexpected weight loss
weakness	Feeling physically weak
Polyphagia	Excessive hunger
Genital thrush	Yeast infection around the genital area
visual blurring	Vision-related disturbances
Itching	Generalized itching
Irritability	Increased irritability
delayed healing	Slow recovery from cuts/wounds
partial paresis	Partial muscle weakness
muscle stiffness	Muscular tightness or inflexibility
Alopecia	Sudden hair loss
Obesity	Clinically overweight condition
class	Diabetes status

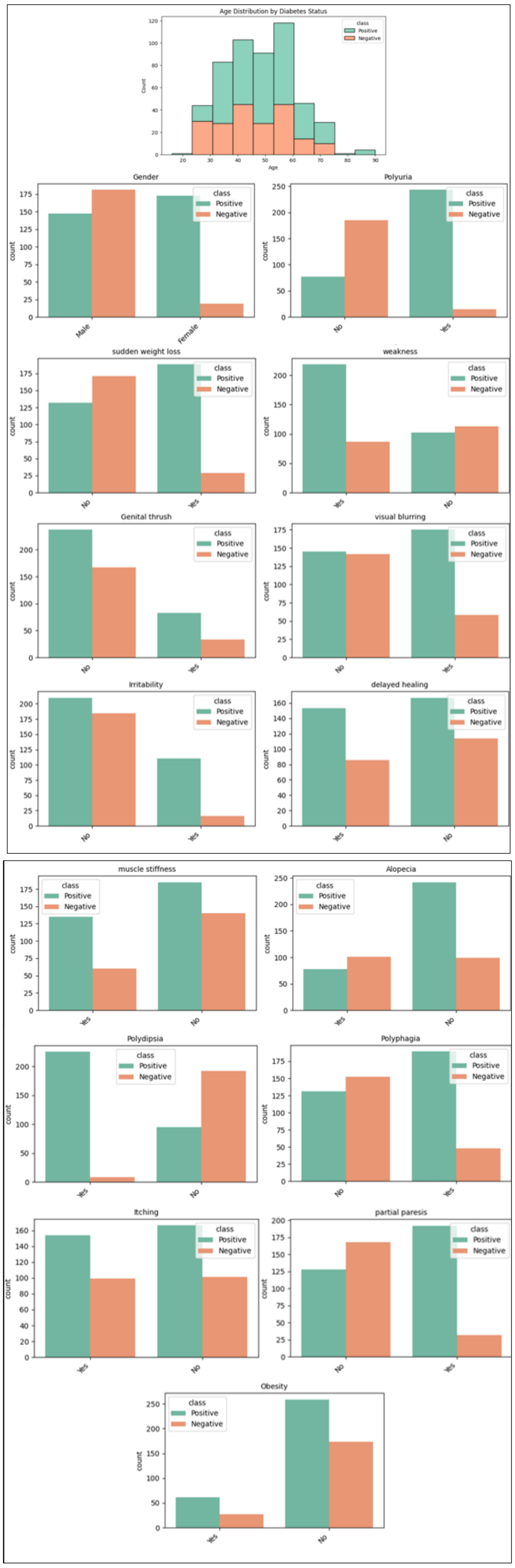


Figure. 4.   Features Distribution in Early-Stage Dataset

imputations. Collectively, these advantages reduce the bias introduced by missing data and contribute to the development of more reliable and generalizable predictive models.

Feature Engineering

To capture complex interactions and domain-specific patterns, the model includes a feature engineering step where additional predictors are derived using both clinical insight and mathematical transformations. Examples include the glucose-to-BMI ratio, blood pressure–glucose product, age–insulin interaction, and a scaled metabolic age (BMI × Age). These engineered features aim to reveal underlying relationships among risk factors that may be overlooked by the original variables, enhancing the model's ability to identify subtle indicators of diabetes risk. As Equation (4), this feature represents the glucose level relative to body mass, potentially highlighting metabolic risks associated with both factors. Equation (5) captures the interaction between blood pressure and glucose, which may jointly influence diabetes risk. Equation (6) models the combined effect of age and insulin levels, which could relate to age-specific insulin resistance. Equation (7) provides a composite measure of metabolic health by scaling the product of BMI and age. Feature construction was guided by clinical knowledge and established risk factor relationships, focusing on biologically meaningful combinations like glucose–BMI and age–insulin interactions. Instead of exhaustively combining variables, we selected features that reflect key physiological mechanisms linked to diabetes. This approach balances interpretability and predictive strength while minimizing overfitting and unnecessary complexity. To identify the most informative and reliable predictors, we applied a hybrid feature selection approach that combines recursive feature elimination with cross-validation, Random Forest importance scores, and Gradient Boosting-based selection. This multi-method strategy assesses feature relevance from different perspectives, reducing dependency on a single technique. Features consistently ranked as important across methods are retained, resulting in a stable, high-quality subset.

$$Glucose\_BMI\_Rtio = \frac{g}{B + \epsilon} \tag{4}$$

$$BP\_Glucose\_Product = BP \times G \tag{5}$$

$$Age\_Insulin\_Interaction = A \times I \tag{6}$$

$$Metabolic\_Age = \frac{B\ \times A}{10} \tag{7}$$

To improve early diabetes detection, several composite features were created by combining related symptom variables into clinically relevant scores. The symptom_count feature measures overall symptom burden by counting the number of 'Yes' responses across all symptom-related items, serving as a general indicator of symptom severity. The hydration_indicator identifies individuals exhibiting both polyuria and polydipsia, which are classic signs of dehydration linked to diabetes. Additional features like infection_risk, neuropathy_indicator, metabolic_stress, and immune_burden aggregate specific symptoms to represent broader physiological conditions such as infection vulnerability, early nerve damage, metabolic imbalance, and immune system strain. These engineered features are designed to capture more complex patterns and interactions, offering stronger predictive value than isolated symptom variables. These composite features are summarized in Table 3.

Hybrid Feature Selection

To select the key variables that contribute most to predictive performance, we employed a hybrid feature selection approach that combines multiple complementary techniques. First, we used Recursive Feature Elimination (RFE) with a Random Forest classifier, which iteratively discards the least important features based on their impact on model accuracy. Let F_{RFE} represent the set of features selected through this process. Next, feature importance scores $I_{RF} = (I_1, I_2, \dots, I_d)$were calculated using the Random Forest model, and the top-ranking features were retained as F_{Imp}. Additionally, we applied model-based selection with a Gradient Boosting classifier, selecting features with non-zero importance values, denoted as F_{GB}. These procedures were repeated across several cross-validation folds to ensure consistency and robustness. The final set of selected features was obtained by taking the union of the features identified by the individual methods as Equation (8). Figures 5–8 present feature importance rankings from Random Forest and Gradient Boosting models across both the PIMA and Early-Stage Diabetes datasets. In the PIMA dataset (Figures 6 and 7), insulin consistently ranks as the most important predictor, followed by engineered features such as the age–insulin interaction and BP × Glucose Product. These results highlight the relevance of both clinical and composite features in diabetes prediction. In the Early-Stage dataset (Figures 8 and 9), importance is distributed across symptom-based features. Metabolic stress and polyuria emerge as top predictors in both models, while other composite features like hydration_indicator, neuropathy_indicator, and symptom_count also contribute meaningfully. Figure 5 illustrates the process of hybrid feature selection.

$$F^* = F_{RFE} \cup F_{Imp} \cup \ F_{GB} \qquad (8)$$

Data Balancing

Clinical datasets used for disease prediction often exhibit class imbalance, where the minority class, typically representing positive cases, is significantly underrepresented. Such imbalance may lead machine learning models to bias their predictions toward the majority class, reducing their ability to correctly identify minority instances. To mitigate this issue, we employed the SMOTETomek technique. SMOTE combined with Tomek Links performs oversampling while also removing overlapping instances between classes to reduce noise [28]. Formally, if N0 and N1 represent the number of majority and minority samples, respectively, SMOTE increases N1 by adding synthetic data, while Tomek link removal reduces both N0 and N1 by eliminating overlapping instances. This combined approach produces a more balanced dataset improves the model's capacity to

Table 3. New Features in Early-Stage Dataset

Feature Name	***Description***
symptom_count	Total number of "Yes" responses for symptoms
hydration_indicator	1 if both polyuria and polydipsia are 1
infection_risk	Sum of itching, genital_thrush, delayed_healing
neuropathy_indicator	Sum of partial_paresis, muscle_stiffness, visual_blurring
metabolic_stress	Sum of polyuria, polydipsia, polyphagia, sudden_weight_loss, weakness
immune_burden	Sum of itching, alopecia, genital_thrush

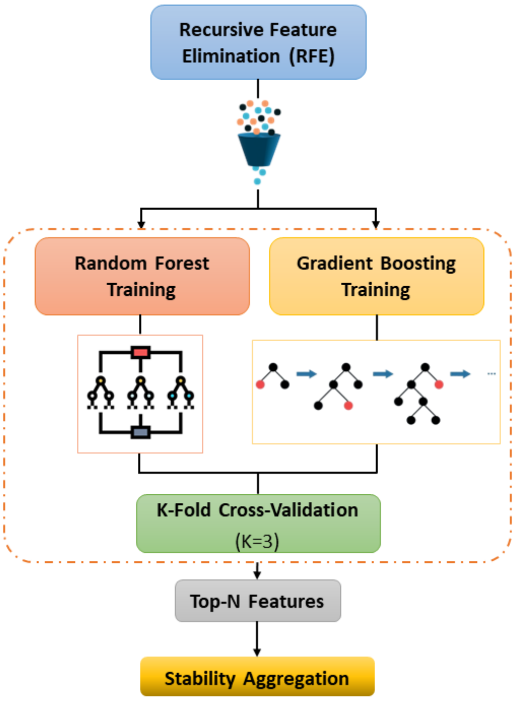

Recursive Feature Elimination (RFE)
Random Forest Training
Gradient Boosting Training
K-Fold Cross-Validation (K=3)
Top-N Features
Stability Aggregation

Figure. 5. Workflow of the hybrid feature selection precess combining multiple evaluation methods

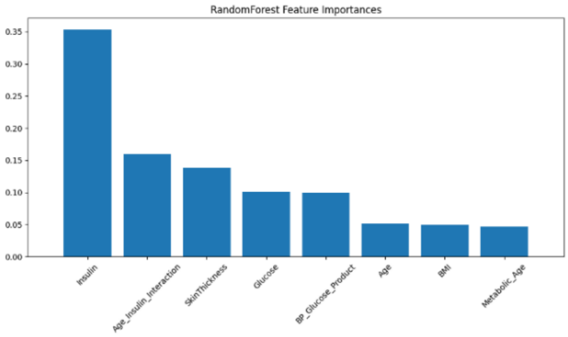


Figure. 6. Feature Importance from Random Forest in PIMA dataset.

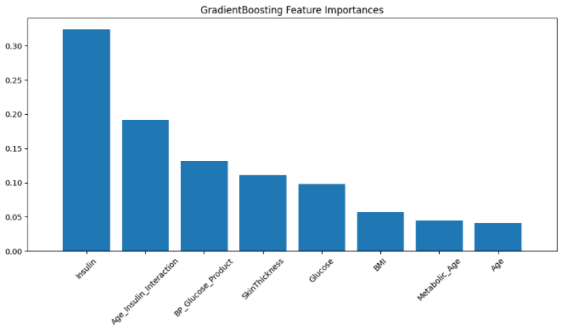


Figure. 7. Feature Importance from Gradient Boosting in PIMA dataset.

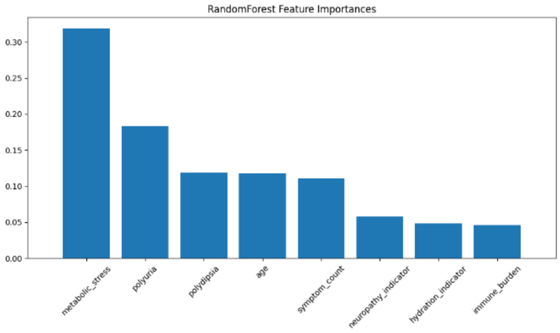


Figure. 8. Feature Importance from Random Forest in Early-Stage dataset.

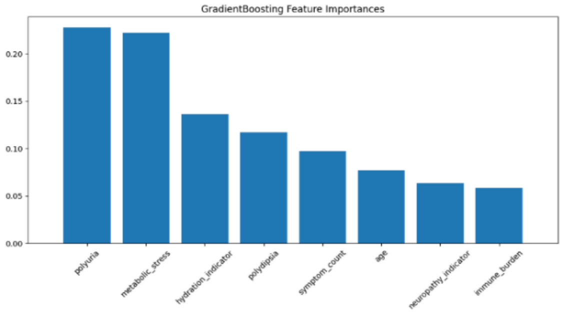


Figure. 9. Feature Importance from Gradient Boosting in Early-Stage dataset.

learn effectively from both classes and improving overall prediction fairness and accuracy.

To provide further transparency regarding dataset characteristics, we analyzed and visualized the class distributions of both datasets used in this study. Figure 10 shows the distribution of diabetic and non-diabetic cases in the PIMA dataset, while Figure 11 illustrates the corresponding distribution in the Early-Stage Diabetes dataset. In the PIMA dataset, we observed 268 diabetic cases compared to 500 non-diabetic cases, indicating a moderate class imbalance where non-diabetic samples dominate. Conversely, in the Early-Stage dataset, 320 diabetic cases and 200 non-diabetic cases were present, resulting in a reverse imbalance where diabetic samples are more frequent.

Feature Scaling

To make sure each feature has an equal influence during the training process, we applied feature scaling to the dataset. Specifically, we used the StandardScaler, which standardizes each feature to have a mean of zero and a standard deviation of one. For each feature x_j, the scaled value $\tilde{x}_j$ is computed as Equation (9). Here, μ_j and σ_j represent the mean and standard deviation of feature x_j across the training data. Normalizing features in this way is critical when variables are measured on different scales, as it prevents those with larger ranges from disproportionately influencing the model. Additionally, feature scaling improves the efficiency of many optimization algorithms and is especially important for distance-based and regularized models. This transformation enhances both the consistency and effectiveness of the machine learning models.

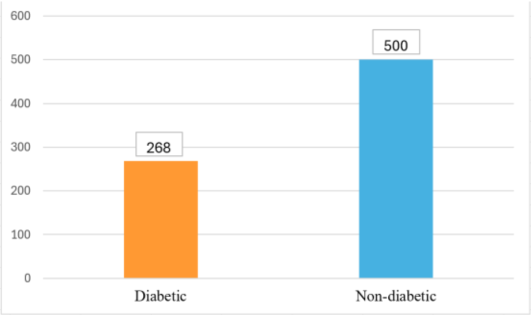


Figure. 10. Distribution of PIMA Dataset

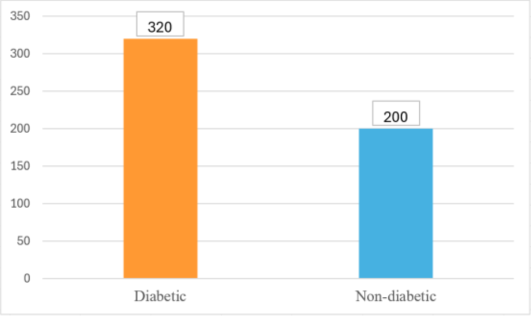


Figure. 11. Distribution of Early-Stage Dataset

$$\tilde{x}_j = \frac{xj - \mu_j}{\sigma j} \qquad (9)$$

4.3. Model Training

For predictive modeling, we employed a supervised learning approach using a balanced and preprocessed dataset with standardized features. The data was split using an 80/20 stratified split was used to divide the data into training and testing sets while preserving the original class distribution. The test set served as an independent benchmark for evaluating model performance using standard metrics: accuracy, precision, recall, F1-score, and ROC-AUC. This setup ensured fair and unbiased performance assessment. In addition to the independent 80/20 split, we also employed stratified k-fold cross-validation during hyperparameter tuning and ensemble weighting. This approach ensured that class proportions were preserved across folds, reduced variance in performance estimates, and provided more reliable evaluation of model generalization. Cross-validation results were further used to determine ensemble weights based on cross-validated ROC-AUC scores. By combining stratified splitting with cross-validation, we strengthened both fairness and robustness in our evaluation pipeline. The dataset was first split into 80% training and 20% testing sets using a stratified split to preserve class distribution. Within the training set, stratified k-fold cross-validation was employed for hyperparameter tuning and ensemble weighting. Finally, the independent test set was used as a benchmark for evaluating the generalization performance of the optimized models.

The computational complexity of the proposed pipeline is influenced by several stages. Preprocessing and feature engineering operate linearly with the dataset size, with complexity $O(n \cdot m)$, where n is the number of samples and m is the number of features. The hybrid feature selection is the most computationally intensive step: recursive feature elimination requires $O(n \cdot m^2)$, Random Forest feature importance scales as $O(n \cdot m \cdot logm)$, and Gradient Boosting selection has a cost of $O(n \cdot m \cdot T)$, where T is the number of boosting iterations. Since these methods are executed under k-fold cross-validation, the overall cost becomes $O(k \cdot (n \cdot m^2 + n \cdot m \cdot logm + n \cdot m \cdot T))$. Balancing with SMOTETomek introduces an additional $O(n \cdot m)$. Model training also contributes significantly: Random Forest training requires $O(n \cdot m \cdot logm \cdot t)$, where t is the number of trees, while Gradient Boosting requires $O(n \cdot m \cdot T)$. Hyperparameter tuning with RandomizedSearchCV multiplies the training cost by the number of configurations N and validation folds k, resulting in $O(N \cdot k \cdot (n \cdot m \cdot logm \cdot t + n \cdot m \cdot T))$. Ensemble prediction and evaluation scale linearly with the number of models, at $O(k \cdot n \cdot m)$. Taken together, the dominant complexity arises from hybrid feature selection and hyperparameter tuning, and the worst-case overall runtime can be summarized as $O(N \cdot k \cdot (n \cdot m^2 + n \cdot m \cdot logm \cdot t + n \cdot m \cdot T))$. While this introduces considerable computational overhead, optimizations such as parallelization, early stopping, and dimensionality reduction can mitigate runtime and improve scalability for larger datasets.

Random Forest and Gradient Boost Machine

We employed two ensemble learning algorithms, Random Forest, which uses bagging, and Gradient Boosting, which relies on boosting, due to their effectiveness in capturing complex feature interactions and enhancing predictive accuracy. Both models were trained separately using the refined feature set obtained from preprocessing. To improve overall robustness and generalization, we combined their outputs through a weighted ensemble approach. The contribution of each model was weighted according to its cross-validated ROC-AUC score, allowing the ensemble to leverage the advantages of both methods.

Hyperparameter Tuning

To optimize model performance and prevent overfitting, hyperparameter tuning was performed using RandomizedSearchCV with stratified cross-validation. This method involved sampling from a predefined search space and selecting the best configuration based on ROC-AUC performance. The final models were retrained on the full training data using the selected parameters to ensure optimal predictive capability on unseen data. Table 4 shows the hyperparameter search space used for RandomizedSearchCV with the RandomForestClassifier. Parameters were randomly sampled to identify the optimal model configuration based on ROC-AUC performance. Table 5 presents the hyperparameter search space for the GradientBoostingClassifier. The tuning process involved randomized sampling from these parameters to maximize predictive accuracy.

4.4. Ensemble Prediction

To enhance prediction accuracy, the outputs of Random Forest and Gradient Boosting models were combined using a weighted averaging method. Each model's prediction was assigned a weight proportional to its cross-validated ROC-AUC score, ensuring that more accurate models contributed more to the final decision. This ensemble approach enhances overall performance by leveraging the strengths of both classifiers.

4.5. Explainability

Explainability is essential in machine learning, especially for applications in healthcare, where trust and transparency are critical [7]. In this study, we used LIME to interpret the predictions made by the

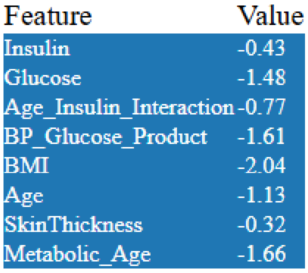
Table 4. Hyperparameter search space in RandomForestClassifier

Hyperparameter	***Values Tried***
n_estimators	[100, 200, 300, 400]
max_depth	[None, 5, 10, 15, 20]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]
max_features	['sqrt', 'log2', 0.5, 0.7]
bootstrap	[True, False]
class_weight	['balanced', None]

Table 5. Hyperparameter search space in GradientBoostingClassifier

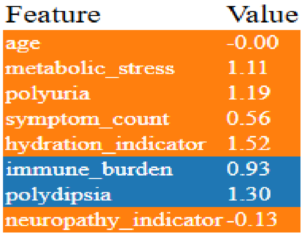
Hyperparameter	***Values Tried***
n_estimators	[100, 200, 300]
learning_rate	[0.01, 0.05, 0.1, 0.2]
max_depth	[3, 4, 5, 6]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]
max_features	['sqrt', 'log2', 0.5]

Random Forest classifier for both the PIMA and Early-Stage Diabetes datasets. Explainable AI methods like LIME help make complex model predictions more interpretable by using simple, local models to approximate the decision-making process [17]. For the PIMA dataset, LIME helped identify key features influencing individual predictions. As shown in Figure 12, features such as BMI, Metabolic Age, BP × Glucose Product, and Glucose showed strong negative contributions to a non-diabetic classification. This indicates that lower values in these features reduced the model's predicted likelihood of diabetes for that instance. The consistent appearance of these features across multiple explanations reinforces their importance and aligns with established clinical risk factors. In the Early-Stage dataset, LIME explanations (Figure 13) highlighted hydration-related symptoms as major predictors. Specifically, polyuria, polydipsia, and the composite hydration_indicator had the strongest positive influence on predicting diabetes. Other contributing features included metabolic stress, immune burden, and symptom count, while features like neuropathy_indicator showed a slight negative influence. These insights emphasize the diagnostic value of grouped symptom features and support the utility of engineered indicators in early detection models. Together, these LIME analyses enhance the explainability of the models and confirm that many of the most influential features align well with medically recognized signs of diabetes, thereby increasing confidence in the model’s clinical relevance.



Feature	Value
Insulin	-0.43
Glucose	-1.48
Age_Insulin_Interaction	-0.77
BP_Glucose_Product	-1.61
BMI	-2.04
Age	-1.13
SkinThickness	-0.32
Metabolic_Age	-1.66

Figure. 12. Feature Values and Their Importance Based on LIME Explanation for a Specific Instance in PIMA dataset



Feature	Value
age	-0.00
metabolic_stress	1.11
polyuria	1.19
symptom_count	0.56
hydration_indicator	1.52
immune_burden	0.93
polydipsia	1.30
neuropathy_indicator	-0.13

Figure. 13. Feature Values and Their Importance Based on LIME Explanation for a Specific Instance in Early-Stage dataset

To complement LIME, we further applied SHAP (SHapley Additive exPlanations), which provides both local (instance-level) and global (dataset-level) interpretability. Unlike LIME, which focuses on localized approximations, SHAP assigns additive contribution values to each feature based on cooperative game theory, enabling a more principled measure of feature importance. For the PIMA dataset (Figures 14 and 15), SHAP highlighted Insulin, SkinThickness, and Age–Insulin Interaction as the most influential predictors, with insulin consistently dominating model output across multiple patients. Similarly, for the Early-Stage dataset (Figures 16 and 17), SHAP emphasized polyuria, metabolic stress, and symptom count as critical drivers of predictions, aligning with both medical knowledge and the engineered features introduced in this study. The combination of LIME and SHAP thus provides a dual perspective: LIME offers intuitive, case-by-case explanations that clinicians can use for individualized decision support, while SHAP ensures a reliable, global understanding of feature contributions across the entire dataset. Together, these complementary approaches significantly enhance the transparency, interpretability, and clinical credibility of our proposed model.

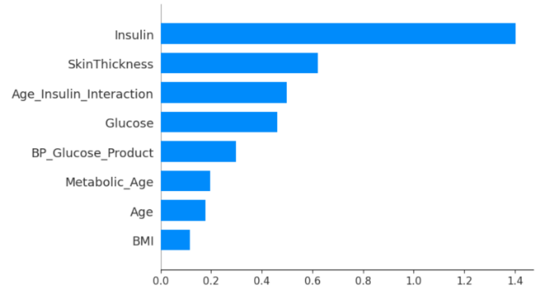


Figure. 14. SHAP bar plot (PIMA dataset): Insulin and SkinThickness dominate feature importance

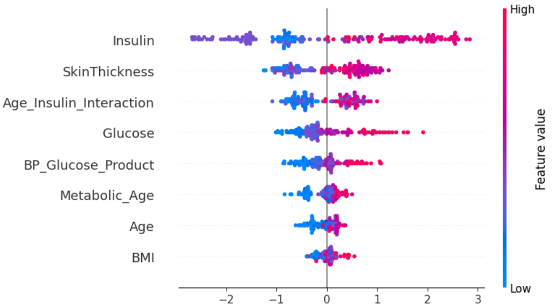


Figure. 15. SHAP beeswarm (PIMA dataset): Insulin, Glucose, and Age–Insulin Interaction drive predictions.

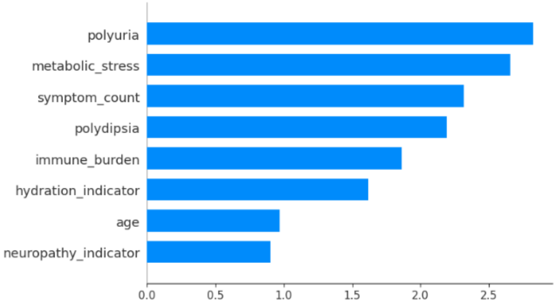


Figure. 16. SHAP bar plot (Early-Stage dataset): Polyuria, metabolic stress, and symptom count are key features.

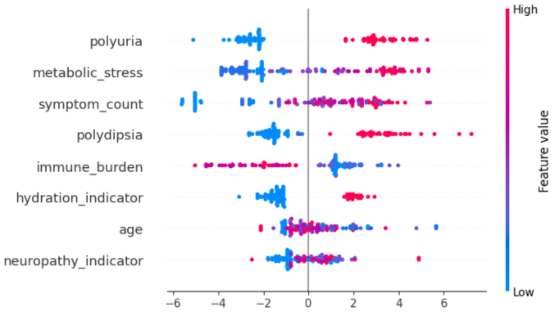


Figure. 17. SHAP beeswarm (Early-Stage dataset): Polyuria and metabolic stress strongly influence diabetes prediction.

5. Results

To quantify how well a model works, several key measurements are used: accuracy, precision, recall, and the F1-score. Accuracy (10) tells us about the overall accuracy by returning the percentage of total cases the model was correct on, it's simple to get a general sense of how well it works. Precision (11) is more concerned with how many of the identified cases that were diabetic were actually correct, and it helps cut down on false alarms or false positives. Recall (12), on the other hand, is interested in the model's ability to detect as many true diabetic cases as it can, attempting to keep false negatives or missed diagnoses to a minimum. F1-score (13) combines precision and recall into one measure by computing their harmonic mean, giving a balanced perspective. As shown in Table 6, the proposed ensemble model achieves strong results on both the PIMA and Early-Stage Diabetes datasets. The model demonstrates balanced performance across accuracy, precision, recall, F1-score, and ROC-AUC, indicating its ability to effectively identify diabetic cases while maintaining low false positives. These results reflect the robustness of the proposed pipeline across different dataset structures. Table 7 compares the model's performance with previous approaches on the PIMA dataset. The proposed method outperforms various models including GRU, XGBoost, and multiple Random Forest implementations. On the Early-Stage dataset, a similar trend is observed in Table 8, where the proposed model surpasses several existing classifiers such as KNN and other Random Forest-based models, further confirming its competitive edge. The ROC curve plots in Figure 18 (PIMA dataset) and Figure 19 (Early-Stage dataset) illustrate the high discriminative power of the ensemble model. In both cases, the ensemble approach performs on par with or slightly better than its individual components, Random Forest and Gradient Boosting, demonstrating improved class separation and reliability. Particularly in the Early-Stage dataset, the ROC curves show near-perfect or perfect performance across models.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \qquad (10)$$

$$Precision = \frac{TP}{TP + FP} \qquad (11)$$

$$Recall = \frac{TP}{TP + FN} \qquad (12)$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (13)$$

In addition to evaluating standard classification metrics, we performed probability calibration to assess how well the predicted probabilities align with

Table 6. Results of Proposed Model on Datasets

Dataset	*Accuracy*	*Precision*	*Recall*	*F1-score*	*ROC-AUC*
PIMA	**93.33**	**95.35**	**93.18**	**93.18**	**97.78**
Early-Stage	**98.44**	**100.00**	**96.88**	**98.41**	**99.89**

Table 7. Comparison Models on PIMA Dataset

Diabetes Prediction Models	*Accuracy*
RFE-GRU [21]	90.70
XGBoost [29]	92.21
RF [30]	80.50
RF [31]	81.70
REMED-T2D [32]	87.80
This Study	**93.33**

Table 8. Comparison Models on Early-Stage Dataset

Diabetes Prediction Models	*Accuracy*
KNN [33]	98.08
RF [25]	97.11
RF [34]	97.91
RF [35]	81.36
This Study	**98.44**

true outcome frequencies, which is particularly relevant for clinical risk scoring.

For the PIMA dataset, Figure 20 shows the ROC curves of the calibrated models, where Random Forest and Gradient Boosting achieved strong ROC-AUC scores of 0.967 and 0.978, respectively, with the ensemble model obtaining 0.976. Importantly, calibration preserved the discriminative ability of the models (AUC remained high). Figure 21 presents the calibration reliability curves, which demonstrate that the predicted probabilities closely follow the diagonal “perfect calibration” line. This confirms that the models provide well-calibrated probability estimates rather than overconfident predictions. Such calibration enhances clinical interpretability by ensuring that predicted risk values can be meaningfully mapped to actual diabetes likelihood.

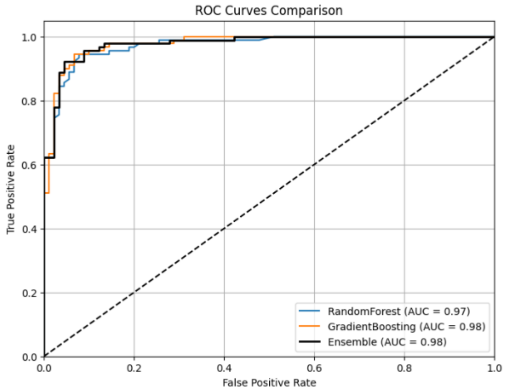


Figure. 18. ROC curve (AUC) comparison of models in PIMA Dataset

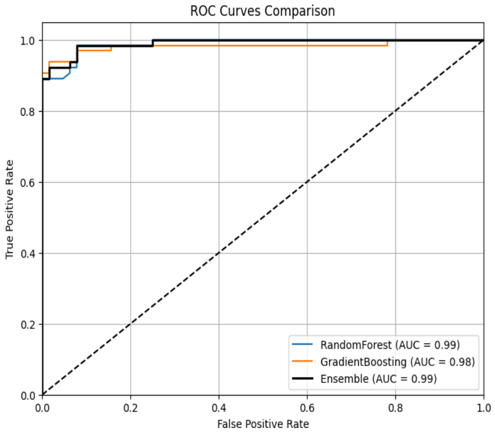


Figure. 19. ROC curve (AUC) comparison of models in Early-Stage Dataset

For the Early-Stage dataset, similar trends were observed (Figures 22 and 23). Both Random Forest and Gradient Boosting produced highly reliable probability outputs, with the ensemble model achieving an ROC-AUC close to 0.998. The corresponding calibration curves show excellent alignment with the ideal calibration line, indicating that probability predictions are not only accurate but also trustworthy for real-world decision-making.

6. Discussion

In medical diagnosis and particularly in disease prediction, several critical factors must be considered. Clinical datasets often present two major challenges: the presence of missing values, which is both common and unavoidable, and class imbalance, where healthy individuals are overrepresented compared to diagnosed patients. These issues can greatly influence the effectiveness of machine learning models if not properly addressed. In this study, careful preprocessing techniques were applied to mitigate these challenges and enhance model performance. To handle missing data, we implemented a class-wise imputation strategy using

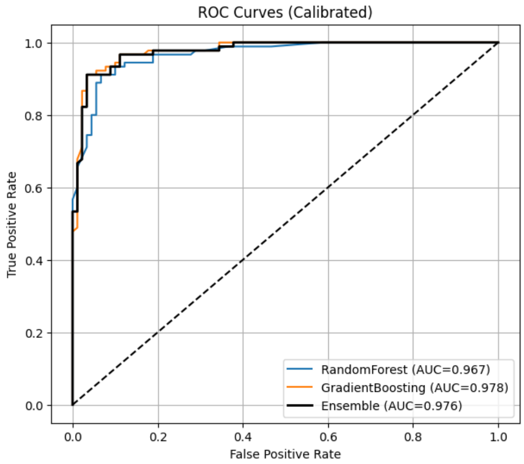


Figure. 20. ROC curves of calibrated models on the PIMA dataset

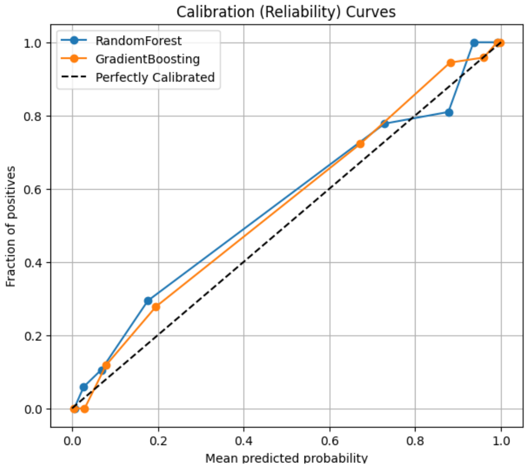


Figure. 21. Calibration (reliability) curves for PIMA dataset

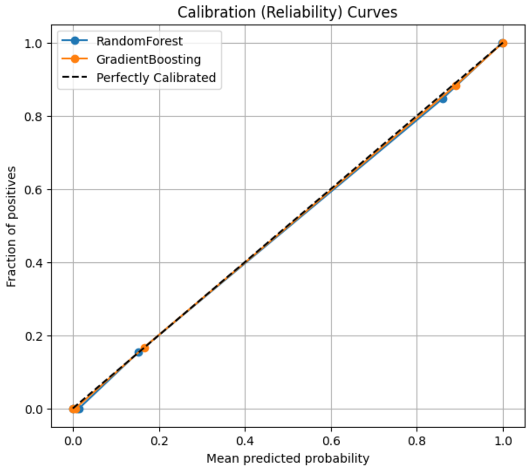


Figure. 22. Calibration (reliability) curves for the Early-Stage dataset

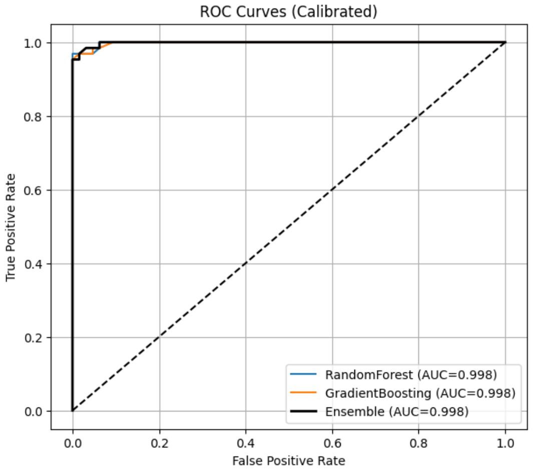


Figure. 23. ROC curves of calibrated models on the Early-Stage dataset

Random Forest Regression, which effectively preserves intra-class feature relationships while reducing the impact of irregularities and extreme values. This method allows the model to generate more accurate estimations based on class-specific patterns. Feature engineering was guided by domain knowledge, with new clinically meaningful features constructed to enhance the model's understanding of symptom interactions and physiological indicators. However, generating features alone is not sufficient; the model must also identify which features are most relevant. To achieve this, we employed recursive feature elimination using Random Forest to iteratively remove less important variables, combined with gradient boosting and cross-validation to ensure stable and robust feature selection. To address class imbalance, we used the SMOTETomek method, which combines oversampling of the minority class with the removal of borderline and overlapping samples. This approach prevents the model from being biased toward the majority class and improves its ability to correctly classify underrepresented positive cases. Ensemble learning was then employed to further enhance prediction accuracy and model stability. By combining Random Forest and Gradient Boosting through weighted averaging, the model benefits from the strengths of both bagging and boosting techniques. This hybrid ensemble approach improves generalizability and reduces the risk of overfitting.

Two datasets were used to evaluate the proposed pipeline. Although the second dataset contained no missing values, it was important to design a model independent of dataset-specific characteristics to ensure broad applicability. The integration of advanced preprocessing, domain-driven feature engineering, robust feature selection, class balancing, and ensemble modeling significantly contributed to the overall reliability and effectiveness of the system in predicting diabetes.

Beyond technical performance, clinical applicability requires considering how healthcare professionals would interact with the system in practice. The integration of explainability tools such as LIME and SHAP provides clinicians with clear insights into which features drive each prediction, but the usefulness of such outputs depends on how they are delivered. A practical extension of this work would be to embed the proposed model into a decision-support dashboard that clinicians can use during consultations. Such a dashboard would not only display predictions (e.g., diabetic vs. non-diabetic) and calibrated probability scores, but also highlight the most influential clinical features for each patient. By presenting results in a transparent, interpretable format, the system could complement physicians' expertise, facilitate trust in model outputs, and enable more informed medical decisions. This bridges the gap between machine learning research and clinical integration, ensuring that the proposed approach can realistically support diagnostic workflows in healthcare environments.

7. Conclusion

In this study, we presented an integrated machine learning model for predicting diabetes, leveraging the strengths of advanced preprocessing, domain-driven feature engineering, hybrid feature selection, and ensemble modeling. By addressing critical challenges such as missing data, class imbalance, and feature relevance, the proposed approach demonstrated strong predictive performance and robustness using clinical datasets. The integration of explainable and statistically sound techniques not only improved model accuracy but also enhanced interpretability, making the framework more suitable for practical use in healthcare environments. Our findings suggest that thoughtful design and integration of machine learning components can significantly improve early detection of diabetes, particularly in resource-constrained or data-limited settings. Although the proposed pipeline performs effectively in identifying diabetes cases, there are several directions for future research. Expanding the study to include external datasets from diverse populations would allow to better assess the model's ability to generalize across different data and reliability in various clinical settings. Additionally, the use of automated or deep learning-based feature engineering techniques could uncover new predictive variables that may not be identified through traditional, domain-driven methods. Incorporating temporal patient data and conducting longitudinal analyses could also improve predictive accuracy and support earlier identification of diabetes risk over time.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' Contributions

The contributions of Ghazaleh Kakavand Teimoory, Mohammad Reza Keyvanpour and Maryam Ghaebi to this paper are as follows:

[GKT]: Conceived the original idea, Designed the methodology, and Wrote the main draft of the manuscript.

[MK]: Supervised the research process, Provided critical feedback, and Contributed to the revision and finalization of the manuscript and Last Edit.

[MG]: Contributed to the coding, Created all figures and diagrams, and Assisted with the preparation and formatting of graphical content.

Conflict of interest

The authors declare that no conflicts of interest exist.

Statement

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with improving the clarity and language of the text. All AI-generated content was carefully reviewed and edited by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

References

[1] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," *Computer Methods and Programs in Biomedicine Update*, vol. 4, p. 100118, Jan. 2023, https://doi.org/10.1016/J.CMPBUP.2023.100118

[2] G. Kakavand Teimoory and M. Keyvanpour, "Elevating Accuracy: Enhanced Feature Selection Methods for Type 2 Diabetes Prediction," *International Journal of Web Research*, vol. 7, no. 2, pp. 37–48, Apr. 2024, https://doi.org/10.22133/IJWR.2024.458872.1218

[3] P. S. Moon, P. A. Bainalwar, S. M. Borkar and S. S. Shambharkar, "Machine Learning approach for Diabetes Prediction using Pima Dataset," in *ACM International Conference Proceeding Series*, *Association for Computing Machinery*, Nov. 2023, pp. 1-9. https://doi.org/10.1145/3647444.3652479

[4] S. A. Tanim, A. R. Aurnob, T. E. Shrestha, M. R. I. Emon, M. F. Mridha and M. S. U. Miah, "Explainable deep learning for diabetes diagnosis with DeepNetX2," *Biomed Signal Process Control*, vol. 99, p. 106902, Jan. 2025, https://doi.org/10.1016/J.BSPC.2024.106902

[5] K. S. Farsana and A. Poulose, "Hybrid Convolutional Neural Networks for PIMA Indians Diabetes Prediction," in *International Conference on Ubiquitous and Future Networks (ICUFN)*, Budapest, Hungary, IEEE Computer Society, 2024, pp. 268–273. https://doi.org/10.1109/ICUFN61752.2024.10624950

[6] M. Zhao *et al.*, "Predictive value of machine learning for the progression of gestational diabetes mellitus to type 2 diabetes: a systematic review and meta-analysis," *BMC Med Inform Decis Mak*, vol. 25, no. 1, p. 18, Dec. 2025, https://doi.org/10.1186/s12911-024-02848-x

[7] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI

techniques," Healthc Technol Lett, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, https://doi.org/10.1049/htl2.12039

[8] N. N. N. Nazirun *et al.*, "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," *IEEE Access*, vol. 12, no. June, pp. 161595–161619, 2024, https://doi.org/10.1109/ACCESS.2024.3432118

[9] H. Lee *et al.*, "Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts from South Korea, Japan, and the UK: a model development and validation study," *EClinicalMedicine*, vol. 80, Feb. 2025, https://doi.org/10.1016/j.eclinm.2025.103069

[10] Q. Sun, X. Cheng, K. Han, Y. Sun, H. Ren and P. Li, "Machine learning-based assessment of diabetes risk: Machine learning-based assessment of diabetes risk," *Applied Intelligence*, vol. 55, no. 2, p. 106, Jan. 2025, https://doi.org/10.1007/s10489-024-05912-1

[11] M. Abroodi, M. R. Keyvanpour and G. K. Teimoory, "Efficient Prediction of Cardiovascular Disease via Extra Tree Feature Selection," *2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Islamic Republic of Iran, 2024, pp. 70–75, https://doi.org/10.1109/ICCKE65377.2024.10874780

[12] H. Lee, M. B. Park and Y. J. Won, "AI Machine Learning-Based Diabetes Prediction in Older Adults in South Korea: Cross-Sectional Analysis," *JMIR Form Res*, vol. 9, no. 1, p. e57874, 2025, https://doi.org/10.2196/57874

[13] M. J. Noh and Y. S. Kim, "Diabetes Prediction Through Linkage of Causal Discovery and Inference Model with Machine Learning Models," *Biomedicines*, vol. 13, no. 1, p. 124, Jan. 2025, https://doi.org/10.3390/biomedicines13010124

[14] M. M. Islam, H. R. Rifat, M. S. Bin Shahid, A. Akhter, M. A. Uddin and K. M. M. Uddin, "Explainable Machine Learning for Efficient Diabetes Prediction Using Hyperparameter Tuning, SHAP Analysis, Partial Dependency, and LIME," *Engineering Reports*, vol. 7, no. 1, p. e13080, Jan. 2025, https://doi.org/10.1002/eng2.13080

[15] F. Mirsharifi and M. R. Keyvanpour, "An EfficientNet-Based Method for Interpretable Early Detection of Alzheimer," *2024 10th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, Shahrood, Islamic Republic of Iran, 2024, pp. 199–204, https://doi.org/10.1109/ICSPIS65223.2024.10931073

[16] M. Taheri, M. R. Keyvanpour and M. S. Mousavi, "Improving Drug-Target Interaction Prediction Using Enhanced Feature Selection," *15th International Conference on Information and Knowledge Technology, (IKT)*, Isfahan, Islamic Republic of Iran, 2024, pp. 157–161, https://doi.org/10.1109/IKT65497.2024.10892664

[17] G. K. Teimoory and M. R. Keyvanpour, "An Explainable Ai Model for Diabetes Prediction Using Random Forest," *2025 11th International Conference on Web Research (ICWR)*, Tehran, Islamic Republic of Iran, Apr. 2025, pp. 264–269, https://doi.org/10.1109/ICWR65219.2025.11006200

[18] A. Agliata, D. Giordano, F. Bardozzo, S. Bottiglieri, A. Facchiano and R. Tagliaferri, "Machine Learning as a Support for the Diagnosis of Type 2 Diabetes," *Int J Mol Sci*, vol. 24, no. 7, p. 6775, 2023, https://doi.org/10.3390/ijms24076775

[19] M. A. Hama Saeed, "Diabetes type 2 classification using machine learning algorithms with up-sampling technique," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, p. 8, 2023, https://doi.org/10.1186/s43067-023-00074-5

[20] H. Zhou, S. Rahman, M. Angelova, C. R. Bruce and C. Karmakar, "A robust and generalized framework in diabetes classification across heterogeneous environments," *Comput Biol Med*, vol. 186, p. 109720, 2025, https://doi.org/10.1016/j.compbiomed.2025.109720

[21] M. Y. Shams, Z. Tarek and A. M. Elshewey, "A novel RFE-GRU model for diabetes classification using PIMA Indian dataset," *Sci Rep*, vol. 15, no. 1, p. 982, 2025, https://doi.org/10.1038/s41598-024-82420-9

[22] S. H. Talukder, S. K. Mondal and R. Bin Sulaiman, "An Efficient Approach for Diabetes Prediction Through Integrated Feature Engineering and Machine Learning," In *2025 4th International Conference on Computing and Information Technology (ICCIT)*, Tabuk, Saudi Arabia, 2025, pp. 451-456, https://doi.org/10.1109/ICCIT63348.2025.10989350

[23] S. S. Bhat, G. A. Ansari and M. D. Ansari, "Performance Analysis of Machine Learning Based On Optimized Feature Selection for Type II Diabetes Mellitus," *Multimed Tools Appl*, vol. 84, pp. 4945–4964, 2024, https://doi.org/10.1007/s11042-024-19000-6

[24] D. Chellappan and H. Rajaguru, "Generalizability of machine learning models for diabetes detection a study with nordic islet transplant and PIMA datasets," *Sci Rep*, vol. 15, no. 1, p. 4479, Dec. 2025, https://doi.org/10.1038/S41598-025-87471-0

[25] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," *Sensors*, vol. 22, no. 14, p. 5247, Jul. 2022, https://doi.org/10.3390/s22145247

[26] M. Karuppasamy, J. M. Rani and K. Poorani, "Metaheuristic Feature Selection for Diabetes Prediction with P-G-S Approach," *Procedia Computer Science*, vol. 252, pp. 165–171, 2025. https://doi.org/10.1016/j.procs.2024.12.018

[27] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," *BMC Med Res Methodol*, vol. 20, p. 199, Jul. 2020, https://doi.org/10.1186/s12874-020-0 1080-1

[28] Y. Jang, "Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study," *Ewha Medical Journal*, vol. 48, no. 2, p. e32, Apr. 2025, https://doi.org/10.12771/emj.2025.00353

[29] M. R. Hossain, M. J. Hossain, M. M. Rahman and M. M. Alam, "Machine Learning Based Prediction and Insights of Diabetes Disease: Pima Indian and Frankfurt Datasets," *Journal of Mechanics of Continua and Mathematical Sciences*, vol. 20, no. 1, pp. 99–114, Jan. 2025, https://doi.org/10.26782/jmcms.2025.01.00007

[30] M. El Sherbiny, M. Abdel Fattah, A. Rabie, A. Taki Eldin and H. Moustafa, "A Diabetes Mellitus Prediction Model Based on Supervised Machine Learning Techniques," *International Journal of Telecommunications*, vol. 5, no. 01, pp. 1-11, 2024. https://doi.org/10.21608/ijt.2025.359269.1083

[31] E. Majeed Hameed, H. Joshi and A. A. A. Ismael, "The Effect of Combining Datasets in Diabetes Prediction Using Ensemble Learning Techniques," *CommIT (Communication and Information Technology) Journal*, vol. 19, no. 1, pp. 129-140, 2025, https://doi.org/10.21512/commit.v19i1.12064

[32] L. T. Phan, R. Rakkiyappan and B. Manavalan, "REMED-T2D: A robust ensemble learning model for early detection of type 2 diabetes using healthcare dataset," *Comput Biol Med*, vol. 187, p. 109771, Mar. 2025, https://doi.org/10.1016/j.compbiomed.2025.109771

[33] O. Julius Adetunji, A. Olusogo Julius, A. Olusola Ayokunle and F. Olawale Ibrahim, "Early Diabetic Risk Prediction using Machine Learning Classification Techniques," *Int. J.*

Innov. Sci. Res. Technol, vol. 9, no. 6, pp. 502-507, 2021. https://www.researchgate.net/publication/369299560

[34] O. O. Oladimeji, A. Oladimeji and O. Oladimeji, "Classification models for likelihood prediction of diabetes at early stage using feature selection," *Applied Computing and Informatics*, vol. 20, no. 3–4, pp. 279–286, Jun. 2024, https://doi.org/10.1108/ACI-01-2021-0022

[35] J. Borges, "Advancing Deep Learning Insights for Identifying Heart Disease in Diabetic Patients: A Data Mining Approach Using Logistic Regression and Random Forests," 2025, https://doi.org/10.2139/SSRN.5091734

Ghazaleh Kakavand Teimoory is currently working toward her master's degree in software engineering, actively engaging in research at the Department of Computer Engineering and Data Mining Lab at Alzahra University, Tehran, Iran. Her research interests are in data mining and its applications, Diabetes Prediction, and diseases analysis and E-health, which contribute significantly to the concept and writing of this article; Tehran, Iran; gh.kakavandteimoory@gmail.com.

Mohammad Reza Keyvanpour is a professor at Alzahra University, Tehran, Iran. His academic journey includes a B.S. in software engineering from Iran University of Science & Technology and his M.S. and Ph.D. in software engineering from Tarbiat Modares University. His research spans information retrieval and data mining and his mainly interest is in E-Health; Tehran; Iran; keyvanpour@alzahra.ac.ir

Maryam Ghaebi received her B.S. degree in Software Engineering from Alzahra University, Tehran, Iran, and is currently pursuing her M.S. degree in software engineering at the Department of Computer Engineering, where she is also an active member of the Data Mining Laboratory, Alzahra University, Tehran, Iran. Her research interests include data mining and its applications in fall detection, human activity recognition, diabetes prediction, and E-health, which have substantially contributed to the conception and preparation of this article; Tehran, Iran; m.ghaebi@student.alzahra.ac.ir