A Sequential Data Preprocessing Pipeline for Diabetes Prediction: A Data Leakage Prevention and Dual-Validation Approach

Ahmed Majid AbdulAbbas
Department of Electrical Engineering, College of Engineering, University of Misan, Amarah, Iraq
ahmedmajed@uomisan.edu.iq (corresponding author)

Rafid Alkanany
Department of Computer Techniques Engineering, Imam Alkadhim University College (IKU), Baghdad, Iraq
rafid.abdalhamed@iku.edu.iq

Yasir Ali Khalid Al-Nuaimi
Department of Electrical Engineering, College of Engineering, University of Misan, Amarah, Iraq
yasseralnaimi6@uomisan.edu.iq

Zahraa Mehssen Agheeb Al-Hamdawee
Department of Electrical Engineering, College of Engineering, University of Misan, Amarah, Iraq
zahraa.mo.eng@uomisan.edu.iq

Received: 19 August 2025 | Revised: 20 September 2025 | Accepted: 6 October 2025

 DOI: https://doi.org/10.48084/etasr.14155

ABSTRACT

Machine learning approaches for diabetes prediction face methodological challenges, including data leakage from preprocessing before data splitting, inconsistent handling of missing values, and class imbalance with varying validation methods. This study presents a systematic approach that prevents data leakage and establishes standardized benchmarks for diabetes prediction. Using the PIMA Indian Diabetes Dataset (768 patients), this study applied a preprocessing pipeline: MICE for missing values (652 missing, 9.43% of data), SMOTE for class balance (500 nondiabetic vs 268 diabetic cases), and z-score normalization for feature scaling. Two feature selection methods identified six important clinical variables: Glucose, Pregnancies, Glucose_BMI, Glucose_Age, BMI, and BloodPressure. Dual validation approaches were employed, single split (80:20) and 5-fold cross-validation, to compare five machine learning algorithms: Random Forest (RF), Multi-Layer Perceptron (MLP), XGBoost, Support Vector Machine (SVM), and Logistic Regression (LR). Experimental results demonstrated that RF achieved the highest accuracy (79.79%) in single split testing, whereas MLP performed best in cross-validation (77.81% accuracy, 84.43% ROC-AUC). All algorithms achieved ROC-AUC scores above 0.80. Cross-validation analysis revealed that RF showed consistent performance across data splits, whereas MLP demonstrated better adaptability to different data conditions.

Keywords-machine learning; diabetes prediction; data preprocessing; cross-validation; data leakage prevention

I. INTRODUCTION

Diabetes mellitus has become a major global health problem, with cases rising rapidly worldwide [1, 2]. More than 800 million adults now have diabetes—a four-fold increase since 1990—which means that 11.1% of adults between 20 and 79 years have diabetes [1, 3]. In 2021, 529 million people had diabetes, and researchers expect this number to reach over 1.3 billion by 2050 [4]. Current diabetes diagnosis methods face problems that make early detection and management difficult [5, 6]. Doctors use plasma glucose tests, such as Fasting Plasma Glucose (FPG), 2-hour glucose during Oral Glucose Tolerance Tests (OGTT), and glycated hemoglobin (HbA1c) levels [7]. These methods have several limitations, as glucose levels vary naturally between people, patients must fast before testing, procedures take considerable time, and hemoglobin variants can affect results [6-9].

Machine learning algorithms offer enhanced capabilities for diabetes detection [10, 11]. Previous studies have demonstrated the effectiveness of these methods for identifying diabetes at early stages, with research publications increasing rapidly since 2010 [10, 12]. Machine learning algorithms analyze large, complex datasets and find patterns that doctors might miss during regular clinical assessment [11, 12]. Multiple machine learning algorithms have been evaluated, including K-Nearest Neighbor (KNN), Decision Trees (DT), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machines, demonstrating consistent performance improvements over traditional diagnostic methods [13]. Combining diverse data sources, such as electronic health records, laboratory parameters, demographic information, and lifestyle factors, enables more comprehensive risk assessment and prediction capabilities [13]. Recent advances include the development of eXplainable Artificial Intelligence (XAI) frameworks, incorporating techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) to enhance model transparency and clinical interpretability [14, 15].

Numerous studies [16-22] have applied machine learning techniques to the prediction of diabetes, with most focusing on the PIMA diabetes dataset. These studies on the PIMA dataset show different accuracy results and use various approaches for data processing and algorithm testing. In [16], machine learning algorithms on the PIMA dataset achieved accuracies between 65-82%. This study compared different algorithms but did not address missing data properly. In addition, SMOTE was used for class balance without testing other methods. In [17], XGBoost with ADASYN was used on data from Bangladeshi female patients, achieving 81% accuracy. This approach was only tested on one population group and did not use cross-validation. Several studies have compared different algorithms. In [18], RF, DT, NB, and LR were compared on PIMA data, with LR achieving 70.5% accuracy and DT reaching 72%, but this study used basic feature engineering and did not optimize parameters well. In [19], six classifiers, including J48, Multi-Layer Perceptron (MLP), and RF, were tested. This study used feature selection, finding that glucose, age, and diabetic pedigree were important variables. Hoeffding Tree performed best, with 75% accuracy using 10-fold cross-validation. In [20], seven algorithms were tested with basic preprocessing, achieving 72-75% accuracy. However, the results were inconsistent and did not include ROC-AUC analysis. In [21], three algorithms, NB, SVM, and DT, were compared, with NB achieving 76.30% accuracy. In [22], two methods were tested on PIMA data, with LR reaching 83% accuracy and DT achieving 82%. Table I compares recent studies on the PIMA dataset with varying methods and results.

Most existing studies have three main problems. First, several studies preprocess the data before splitting it into training and testing sets. This can cause data leakage where information from test data affects training, making accuracy scores higher than they should be. Second, studies handle missing values and class imbalance inconsistently. Some ignore missing data [16], others use basic imputation methods [18], and many apply only one balancing technique without systematic comparison [17]. Third, validation methods vary between studies, since some use single splits [16, 17], whereas others use cross-validation [19], making a fair comparison difficult.

TABLE I. COMPARISON OF RECENT STUDIES ON THE PIMA DATASET

Study	Method	Accuracy	Validation	Preprocessing
[16]	Multiple ML	65-82%	Single split	Poor missing data handling
[17]	XGBoost/ADASYN	81%	No CV	ADASYN only
[18]	LR/DT	70.5-72%	Basic	Basic feature engineering
[19]	Six classifiers	75%	10-fold CV	Feature selection
[20]	Seven algorithms	72-75%	Basic	Basic preprocessing
[21]	NB/SVM/DT	76.30%	Not specified	Not specified
[22]	LR/DT	83%/82%	Not specified	Preprocessing method unclear

Different preprocessing approaches make comparison challenging. Studies use varying methods to handle missing values, scale features, and balance classes. When studies report different accuracies, it remains unclear whether improvements stem from better algorithms or superior data preparation. Most studies only report accuracy, omitting other important measures such as ROC-AUC or precision-recall.

This study aimed to address these methodological limitations by implementing a systematic preprocessing pipeline after data splitting to prevent data leakage. The proposed method applies MICE for missing value imputation, SMOTE for class balancing, and standardization for feature scaling. Two feature selection approaches are compared, and dual validation strategies are employed: single split evaluation and 5-fold cross-validation. This systematic approach compares algorithms under identical preprocessing conditions, providing reliable benchmarks for assessing the performance of diabetes prediction.

II. DATASET ANALYSIS AND PREPROCESSING

The PIMA diabetes dataset contains several data quality issues, particularly disguised missing values represented as zeros in clinical measurements [23]. The dataset has 652 missing values (9.43% of the total data). Missing values occur mainly in three features: Insulin (374 missing, 48.7%), SkinThickness (227 missing, 29.6%), and BloodPressure (35 missing, 4.6%). These zero values represent missing data, not actual measurements [23]. Missing data occurs frequently in clinical settings due to measurement difficulties [24]. Insulin measurements have more missing data because they are more invasive and costly than simple measurements such as age or pregnancy history.

The dataset also shows class imbalance with 500 nondiabetic cases versus 268 diabetic cases (65% negative versus 35% positive). This imbalance reflects the natural disease prevalence in screening populations [25]. This class imbalance can bias machine learning algorithms toward the majority class and reduce sensitivity for detecting diabetic patients [26]. Correlation analysis shows relationships between features, as shown in Figure 1. Strong correlations include Age-Pregnancies (0.54), SkinThickness-BMI (0.53), and

Insulin-SkinThickness (0.44). These patterns match findings from previous research [27-29]. The Glucose-Outcome correlation (0.47) is the strongest relationship, confirming glucose as the main factor for diabetes diagnosis [30]. Blood pressure shows weak correlations with other features (below 0.3), suggesting it acts as an independent factor [31]. Understanding these relationships helps with feature selection in medical prediction [32].

The preprocessing follows a specific sequence: missing value imputation, class balancing, and feature standardization. This order ensures complete data availability before other transformations and prevents bias during processing.

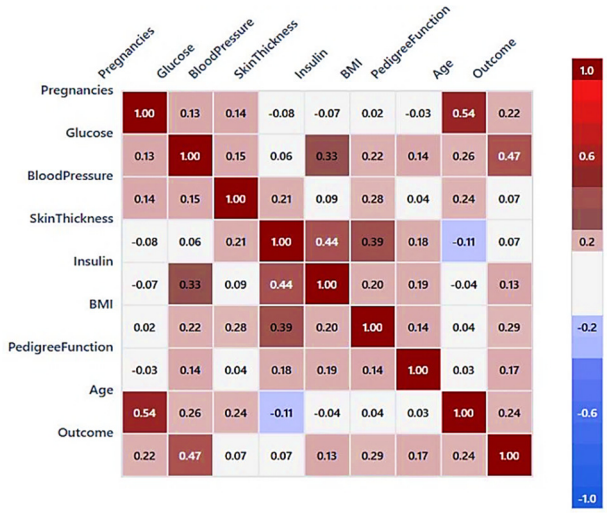

Fig. 1. Pearson correlation matrix heatmap for PIMA Indian Diabetes Dataset features.

A. *Multiple Imputation by Chained Equations (MICE)*

MICE handles missing data by modeling each variable as a function of other variables using regression equations [33]. This algorithm looks at each variable with missing values and predicts them based on all other variables in the dataset as:

$$Y_j^{(t+1)} = \beta_0 + \beta_1 X_1^{(t)} + \beta_2 X_2^{(t)} + \cdots + \beta_p X_p^{(t)} + \varepsilon \quad (1)$$

where Y_j^{t+1} represents the imputed values for variable j at iteration $t+1$, and $X_1^t, X_2^t, \ldots, X_p^t$ denote the current values of predictor variables. This process examines each variable with missing data one by one, updating the predicted values based on the latest estimates from other variables. The random component ε accounts for uncertainty in the predictions.

MICE successfully filled all missing values in the diabetes dataset, handling missing values in the five clinical features, particularly in Insulin, SkinThickness, and BloodPressure measurements. This method preserved the original data relationships while providing complete data for machine learning model training.

B. *Synthetic Minority Oversampling Technique (SMOTE)*

SMOTE addresses class imbalance by generating synthetic samples using a mathematical interpolation approach [34]. This algorithm identifies k-nearest neighbors for each minority class instance (X_i) and creates synthetic examples along the line segments connecting these neighbors using:

$$X_{new} = X_i + \lambda \times (X_{zi} - X_i), \text{ where } \lambda \in [0,1] \quad (2)$$

where X_{zi} represents a randomly selected neighbor and λ is a random value between zero and one that determines the exact position of the synthetic sample along the connecting line. This interpolation-based approach generates more diverse and strategically positioned training examples compared to random oversampling, as the random selection of both neighbors and interpolation coefficients ensures variability while maintaining the underlying data distribution characteristics, which may help models perform better on new data and avoid overfitting.

SMOTE was used to handle the class imbalance in the diabetes dataset. As illustrated in Figure 2, the original dataset showed a significant class imbalance with a 1.87:1 ratio between nondiabetic (500 samples, 65.1%) and diabetic cases (268 samples, 34.9%), which could bias machine learning algorithms toward the majority class and reduce sensitivity in minority class detection. SMOTE successfully addressed this imbalance by generating 232 synthetic minority class samples, transforming the dataset from 768 to 1000 total samples while achieving perfect class balance (1:1 ratio). This oversampling approach offers equal learning opportunities for both classes during model training. The balanced dataset resulting from SMOTE allows subsequent machine learning algorithms to reach improved accuracy in identifying both diabetic and nondiabetic patients, eliminating the bias toward majority class prediction that typically affects imbalanced medical datasets.

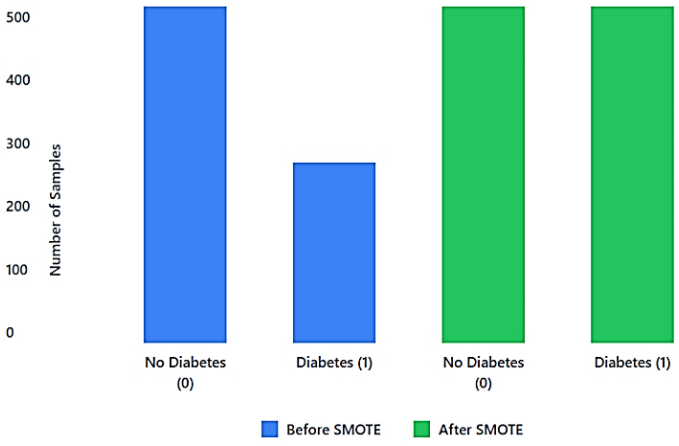

Fig. 2. Results after SMOTE - Class balance transformation.

C. *Feature Standardization Analysis*

Feature standardization (z-score normalization) transforms variables to have a mean of zero and unit variance, addressing the scale differences that can bias distance-based algorithms. The transformation follows the formula:

$$Z = (X - \mu) / \sigma \quad (3)$$

where Z represents the standardized value, X is the original feature value, μ is the mean, and σ is the standard deviation. This process ensures that all features contribute equally to distance calculations, preventing variables with larger scales (such as Insulin, ranging 0-800) from dominating those with smaller scales (such as DiabetesPedigreeFunction, ranging 0-2.5). This normalization helps machine learning algorithms that use distance calculations, including SVM, Neural Networks (NNs), and KNN [35].

The preprocessing follows a specific sequence: missing value imputation, class balancing, and then feature standardization [36]. This order ensures complete data availability before other transformations and allows SMOTE to use original feature scales for accurate nearest neighbor identification during synthetic sample generation.

III. FEATURE SELECTION METHODS

Feature selection is important in medical machine learning, especially for diabetes prediction. The goal is to achieve good accuracy while keeping models simple enough for doctors to understand and use. Medical datasets often have many variables related to each other, which can cause overfitting, slow processing, and poor performance on new patients [32]. For diabetes prediction, the selected features should improve accuracy and represent biological factors that help explain disease development. This requires selecting methods that can identify both direct and indirect relationships between variables in diabetes development [37].

This study uses two feature selection methods: Mutual Information (MI)-based selection to find nonlinear relationships between variables, and Recursive Feature Elimination (RFE) to remove less important features step by step.

A. Mutual Information (MI)-Based Feature Selection

MI scores measure how much each feature relates to the target variable. This method can find both linear and nonlinear relationships in data. For feature X and target Y, MI is calculated as:

$$I(X;Y) = \sum\sum p(x,y) \log_2(p(x,y)/(p(x)p(y))) \quad (4)$$

where $I(X;Y)$ is the MI between feature X and target Y, $p(x,y)$ is the joint probability distribution $p(x)$, $p(y)$ is the marginal probability distribution, and $\log_2$ is the binary logarithm for information measured in bits. Higher MI scores indicate greater predictive value. When MI equals zero, variables are independent. The SelectKBest algorithm with MI classification ranked all features in the diabetes dataset. Glucose achieved the highest score (0.175), followed by Pregnancies (0.140). Engineered features also performed well, such as Glucose_BMI (0.120) and Glucose_Age (0.110), demonstrating that feature combinations enhance predictive capability. The remaining features scored lower: BMI (0.095), BloodPressure (0.090), SkinThickness (0.088), Age (0.087), Insulin_BMI (0.085), and Insulin (0.075). These rankings confirm glucose levels and pregnancy history as primary diabetes prediction factors.

B. Recursive Feature Elimination (RFE)

RFE is an advanced wrapper-based feature selection approach that evaluates feature subsets through iterative model training. Unlike filter methods such as MI, RFE considers feature interactions and dependencies within the context of the specific learning algorithm, making it effective for complex predictive tasks [38]. The RFE optimization process can be defined as:

$$RFE(D,k,M) = argmin\_\{S \subseteq F, |S| = k\}\, L(M(D_s)) \quad (5)$$

where D represents the training dataset, k denotes the desired number of features, F is the complete feature set, S is the selected feature subset, M is the base machine learning model, and L is the loss function applied to the dataset restricted to features in subset S.

XGBoost served as the base model for RFE. XGBoost calculates feature importance by measuring contributions across all trees [39]. The algorithm iteratively removes the least important features until reaching the desired number. RFE confirmed the same top-6 features identified by MI with identical rankings: Glucose (0.175), Pregnancies (0.140), Glucose_BMI (0.120), Glucose_Age (0.110), BMI (0.095), and BloodPressure (0.090). This consistency across different selection methods validates feature importance. The engineered features (Glucose_BMI and Glucose_Age) scored higher than individual Age, showing that combining features can improve prediction. These selected features, detailed in Table II, match known diabetes risk factors from medical research:

- Glucose is the main diagnostic measure that shows pancreatic function and insulin effectiveness.
- Pregnancies: The history of gestational diabetes increases Type 2 diabetes risk by 2-7 times.
- Glucose_BMI shows the relationship between blood sugar control and body weight.
- Glucose_Age reflects how glucose tolerance decreases with age.

BMI and Blood Pressure help identify metabolic syndrome and insulin resistance.

IV. RESULT AND DISCUSSION

This study evaluated five machine learning algorithms in diabetes prediction. Algorithm selection depends on several factors: interpretability for doctors, performance on unbalanced datasets, ability to find feature interactions, and handling outliers. The algorithms tested were LR, RF, SVM with RBF kernel, XGBoost, and MLP. The five diabetes prediction models were compared using a systematic evaluation, beginning with a confusion matrix. The confusion matrix quantifies four types of predictions the models can produce:

- True Positives (TP): Correctly identified diabetic patients.
- True Negatives (TN): Correctly identified nondiabetic patients.

- False Positives (FP): Nondiabetic patients incorrectly labeled as diabetic.
- False Negatives (FN): Diabetic patients missed by the model.

From these four basic counts, important performance metrics can be calculated to determine different aspects of how well each model works. Accuracy measures the overall classification performance by calculating the proportion of correct predictions:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (6)$$

Precision quantifies the reliability of positive predictions by calculating the proportion of true positive predictions among all positive classifications:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

Recall evaluates the model's sensitivity by calculating the proportion of actual positive cases that are correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (8)$$

F1-score provides a balanced measure that combines precision and recall into a single number:

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

TABLE II. FEATURE SELECTION RESULTS AND CLINICAL INTERPRETATION

MI+RFE rank	Feature name	MI score	Clinical significance	Biological rationale
1	Glucose	0.175	Direct measure of blood sugar levels. Primary diagnostic criterion for diabetes (fasting glucose ≥ 7.0 millimoles per liter)	Reflects pancreatic beta-cell function and insulin effectiveness. Core pathophysiology of diabetes
2	Pregnancies	0.140	A history of gestational diabetes increases Type 2 diabetes risk by 2-7 times	Pregnancy causes insulin resistance. Multiple pregnancies stress the pancreatic beta-cells
3	Glucose_BMI	0.120	Combined glucose and obesity indicator. Enhances diabetes risk prediction accuracy	Interaction between blood sugar and body weight. Obesity worsens glucose control
4	Glucose_Age	0.110	Age-adjusted glucose levels. Accounts for the natural decline in glucose tolerance with aging	Beta-cell function decreases with age. Older adults have reduced glucose processing ability
5	BMI	0.095	BMI measures obesity. Each unit increase above 22 kg/m^2 raises diabetes risk by 30%	Excess fat tissue produces hormones that interfere with insulin function
6	BloodPressure	0.090	High blood pressure commonly occurs with diabetes. Part of metabolic syndrome	Insulin resistance affects blood vessel function. Shared pathway with diabetes development

A. Initial Performance Assessment (Single Split Evaluation)

The initial performance assessment employed a straightforward method: the diabetes dataset was randomly split into two parts—80% for training and 20% for testing the models. This approach serves as a preliminary screening to identify which algorithms show the most promise before moving to more rigorous methods. Each of the five models underwent the same test: they were trained on 614 patient records and then asked to predict diabetes status for the remaining, previously unseen, 154 patients. The preprocessing pipeline ensured that all models received the same high-quality input data, complete records thanks to MICE imputation, and balanced classes through SMOTE and standardized features for fair comparison. The detailed analysis of these results reveals important insights into each algorithm's performance characteristics and optimization parameters.

- RF was the superior performer after hyperparameter optimization (*n_estimators*=100, *min_samples_split*=5, *max_depth*=10, *min_samples_leaf*=2), obtaining the highest accuracy (79.79%) and F1-score (79.72%). This ensemble method combined 100 decision trees with bootstrap aggregating, demonstrating exceptional balance between precision (76.51%) and recall (83.21%), effectively capturing complex nonlinear relationships while maintaining robust generalization capability.
- MLP achieved the second-highest accuracy at 76.66%, demonstrating the effectiveness of a deep neural network architecture with hidden layers (100, 50), ReLU activation functions, and Adam optimizer (learning rate=0.001). MLP achieved the best precision (77.78%), indicating excellent ability to minimize FP predictions, although with moderate recall (71.53%).
- SVM with RBF kernel (*C*=1.0, *γ*='scale') demonstrated the highest sensitivity (84.67% recall), making it exceptionally effective at detecting true positive diabetes cases through kernel transformation for optimal hyperplane identification.
- XGBoost with optimized parameters (*max_depth*=6, *learning_rate*=0.1, *n_estimators*=100, *subsample*=0.8) achieved competitive performance (76.31% accuracy), demonstrating the effectiveness of gradient boosting's sequential error correction mechanism.
- LR served as the interpretable baseline (70.38% accuracy), providing a linear decision boundary interpretability that is valuable for clinical insights. Feature importance analysis revealed Glucose_Age and Glucose_BMI as the most predictive engineered features.

Table III presents the complete comparative analysis of this initial evaluation. Different algorithms perform better in different areas, and algorithm choice depends on whether accuracy, sensitivity, or precision is most important. RF showed the best overall performance in this initial evaluation.

Confusion matrices show where each model makes specific misclassifications. In medical diagnosis, different errors have different consequences. Missing a diabetic patient (FN) delays treatment, whereas wrongly diagnosing a healthy person as

diabetic (FP) causes unnecessary tests. Figure 3 shows these error patterns for all models. The experimental results of the initial single split evaluation showed that RF achieved the highest accuracy (79.79%), followed by MLP (76.66%). However, these single split results require further validation through cross-validation.

TABLE III. PERFORMANCE COMPARISON FOR DIABETES DIAGNOSIS (SINGLE SPLIT: 80:20)

Model	Accuracy	Precision	Recall	F1-Score
RF	0.7979	0.7651	0.8321	0.7972
MLP	0.7666	0.7778	0.7153	0.7452
SVM-RBF	0.7631	0.7117	0.8467	0.7733
XGBoost	0.7631	0.7482	0.7591	0.7536
LR	0.7038	0.7000	0.6642	0.6816

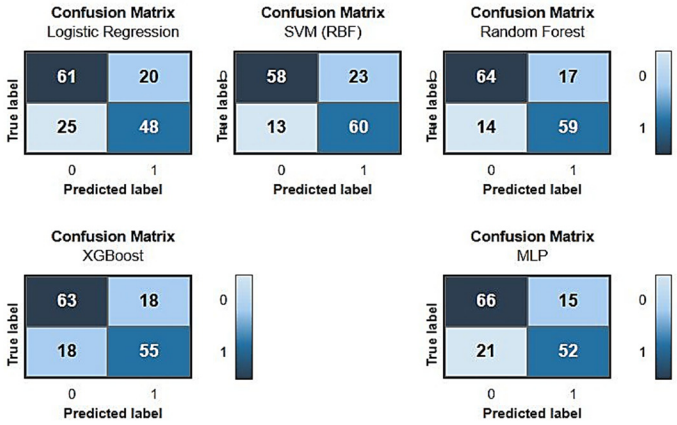

Fig. 3. Confusion matrices for the five machine learning algorithms (single-split validation).

ROC (Receiver Operating Characteristic) curves measure how well models distinguish between nondiabetic and diabetic cases at different probability thresholds, plotting the TP rate against the FP rate. The Area Under the Curve (AUC) summarizes this performance, with values near 1.0 indicating perfect classification and near 0.5 indicating random guessing. Figure 4 presents a ROC curve comparison for all five evaluated models. RF obtained the highest AUC of 0.851.

B. Cross-Validation Assessment

This study applied a 5-fold stratified cross-validation to the training dataset (614 patients), keeping 154 patients separate for final testing. The training data was divided into five groups of approximately 123 patients each, maintaining the same proportion of diabetic and nondiabetic cases. Each model was trained five times, with four groups used for training and one for validation in each round. Table IV shows the cross-validation results with mean performance and standard deviations. Cross-validation results showed important changes from single split testing. MLP achieved the best mean accuracy (77.81%), whereas RF scored 75.56%, XGBoost reached 76.31%, SVM scored 75.86%, and LR achieved 68.22%. The standard deviations show model stability: RF had the lowest variation (±0.0066), whereas MLP had slightly more variation (±0.0250) across different data splits. Figure 5 shows these results with error bars representing standard deviations.

TABLE IV. CROSS-VALIDATION RESULTS [(MEAN ± STANDARD DEVIATION (STD)]

Model	Accuracy	Precision	Recall	F1	ROC_AUC
Lr	0.6822 ± 0.0463	0.6926 ± 0.0751	0.6151 ± 0.0214	0.6494 ± 0.0343	0.7800 ± 0.0451
SVM (RBF)	0.7586 ± 0.0272	0.7078 ± 0.0264	0.8388 ± 0.0449	0.7672 ± 0.0284	0.8056 ± 0.0318
RF	0.7556 ± 0.0066	0.7202 ± 0.0228	0.7983 ± 0.0342	0.7562 ± 0.0052	0.8372 ± 0.0172
XGBoost	0.7631 ± 0.0190	0.7356 ± 0.0181	0.7858 ± 0.0672	0.7580 ± 0.0286	0.8312 ± 0.0161
MLP	0.7781 ± 0.0250	0.7551 ± 0.0323	0.7920 ± 0.0351	0.7724 ± 0.0240	0.8443 ± 0.0217

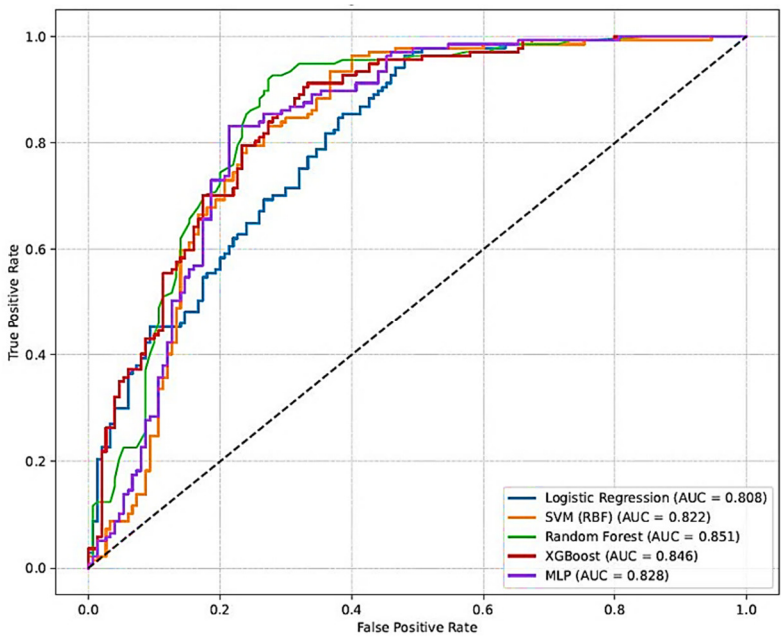

Fig. 4. ROC curve comparison of models (single split assessment).

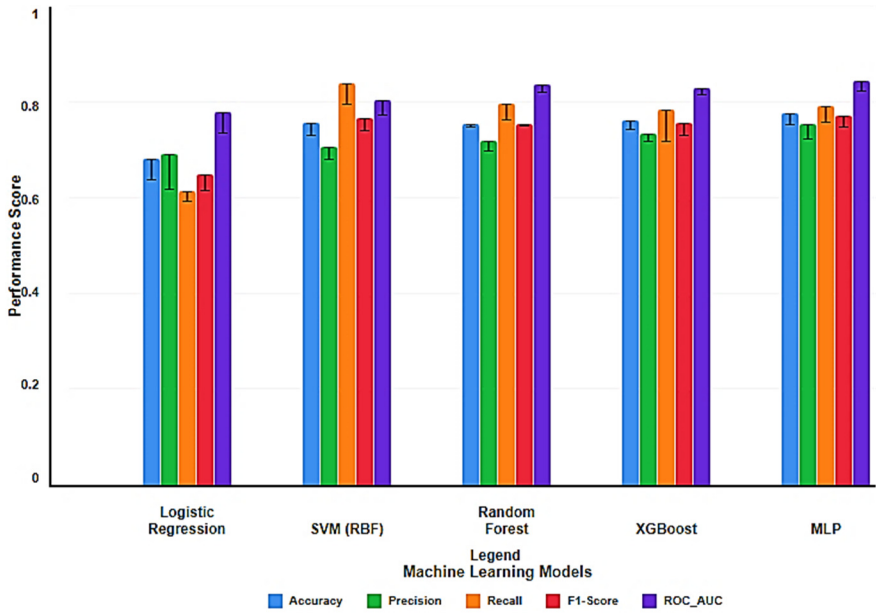

Fig. 5. Five-fold cross-validation results for diabetes prediction models.

V. CONCLUSIONS AND RECOMMENDATIONS

Research on diabetes prediction faces several problems that affect model performance in clinical settings. Many studies preprocess data before splitting them into training and testing sets, leading to data leakage and inflated accuracy scores. Other studies handle missing values and class imbalance differently, using inconsistent validation methods. These problems make the reported results unreliable for medical use. This study addresses these problems through systematic preprocessing and dual validation methods, making three important contributions to the prediction of diabetes. First, data splitting before preprocessing prevents data leakage that compromises model validity. MICE imputation, SMOTE balancing, and feature standardization are applied only after data separation. This approach ensures that test data never influences training, providing reliable performance estimates for clinical use. Second, algorithm testing using both single split and cross-validation reveals different performance characteristics. The single split results showed that the RF achieved the highest accuracy (79.79%), whereas cross-validation demonstrated that MLP performed best overall (77.81% accuracy, 84.43% ROC-AUC). RF shows consistent performance across data splits, whereas MLP adapts better to varying conditions. Third, feature selection identified six important clinical variables while demonstrating that engineered features outperform individual components. Both MI and RFE selected identical features: Glucose, Pregnancies, Glucose_BMI, Glucose_Age, BMI, and BloodPressure. Engineered features (Glucose_BMI, Glucose_Age) scored higher than individual components, confirming that feature combinations capture important diabetes risk relationships. All algorithms achieved ROC-AUC scores above 0.80, providing reliable benchmarks for healthcare providers evaluating diabetes prediction systems.

This study establishes methodological standards for medical machine learning research. The preprocessing sequence and dual validation strategy should become standard practice for diabetes prediction studies. Future research can build on this foundation by testing the proposed approach in different populations, integrating additional clinical variables, and exploring ensemble methods that combine the strengths of multiple algorithms.

REFERENCES

[1] "Urgent action needed as global diabetes cases increase four-fold over past decades," *World Health Organization*. https://www.who.int/news/item/13-11-2024-urgent-action-needed-as-global-diabetes-cases-increase-four-fold-over-past-decades.

[2] Md. J. Hossain, Md. Al-Mamun, and Md. R. Islam, "Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused," *Health Science Reports*, vol. 7, no. 3, Mar. 2024, Art. no. e2004, https://doi.org/10.1002/hsr2.2004.

[3] "Facts & figures," *International Diabetes Federation*. https://idf.org/about-diabetes/diabetes-facts-figures/.

[4] K. L. Ong *et al.*, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021," *The Lancet*, vol. 402, no. 10397, pp. 203–234, Jul. 2023, https://doi.org/10.1016/S0140-6736(23)01301-6.

[5] "Diagnosis and Classification of Diabetes: *Standards of Care in Diabetes—2024*," *Diabetes Care*, vol. 47, no. s1, pp. S20–S42, Jan. 2024, https://doi.org/10.2337/dc24-S002.

[6] N. Hussain, "Implications of using HBA1C as a diagnostic marker for diabetes," *Diabetology International*, vol. 7, no. 1, pp. 18–24, Nov. 2015, https://doi.org/10.1007/s13340-015-0244-9.

[7] S. I. Sherwani, H. A. Khan, A. Ekhzaimy, A. Masood, and M. K. Sakharkar, "Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients," *Biomarker Insights*, vol. 11, Jan. 2016, Art. no. BMI.S38440, https://doi.org/10.4137/BMI.S38440.

[8] "Spotlight on limitations of the HbA1c test," *ACP Diabetes Monthly*. https://diabetes.acponline.org/archives/2024/04/12/5.htm.

[9] O. Schnell, J. B. Crocker, and J. Weng, "Impact of HbA1c Testing at Point of Care on Diabetes Management," *Journal of Diabetes Science and Technology*, vol. 11, no. 3, pp. 611–617, May 2017, https://doi.org/10.1177/1932296816678263.

[10] M. Kiran, Y. Xie, N. Anjum, G. Ball, B. Pierscionek, and D. Russell, "Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature

analysis," *Frontiers in Digital Health*, vol. 7, Mar. 2025, Art. no. 1557467, https://doi.org/10.3389/fdgth.2025.1557467.

[11] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 24153–24185, Aug. 2023, https://doi.org/10.1007/s11042-023-16407-5.

[12] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, Dec. 2022, Art. no. 196, https://doi.org/10.1186/s13098-022-00969-9.

[13] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, Jun. 2022, Art. no. 100071, https://doi.org/10.1016/j.dajour.2022.100071.

[14] 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," *Biomedical Signal Processing and Control*, vol. 99, Jan. 2025, Art. no. 106902, https://doi.org/10.1016/j.bspc.2024.106902.

[15] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App," *International Journal of Intelligent Systems*, vol. 2024, pp. 1–13, Jan. 2024, https://doi.org/10.1155/2024/6688934.

[16] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," *Diagnostics*, vol. 13, no. 14, Jul. 2023, Art. no. 2383, https://doi.org/10.3390/diagnostics13142383.

[17] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, https://doi.org/10.1049/htl2.12039.

[18] A. Ahmed *et al.*, "Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset," *Healthcare*, vol. 13, no. 1, Dec. 2024, Art. no. 37, https://doi.org/10.3390/healthcare13010037.

[19] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," *Procedia Computer Science*, vol. 112, pp. 2519–2528, Jan. 2017, https://doi.org/10.1016/j.procs.2017.08.193.

[20] N. Ahmed *et al.*, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, Jun. 2021, https://doi.org/10.1016/j.ijcce.2021.12.001.

[21] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018, https://doi.org/10.1016/j.procs.2018.05.122.

[22] R. Barakeh, "Leveraging Machine Learning for Precise Prediction of Type 2 Diabetes," *Diabetes*, vol. 73, no. s1, Jun. 2024, https://doi.org/10.2337/db24-59-PUB.

[23] V. Jain, S. Shukla, and N. Khare, "Analysis of various data imputation techniques for diabetes classification on PIMA dataset," in *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, Feb. 2024, pp. 1–6, https://doi.org/10.1109/SCEECS61402.2024.10482050.

[24] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, Dec. 1976, https://doi.org/10.1093/biomet/63.3.581.

[25] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, Sep. 2002, https://doi.org/10.3233/IDA-2002-6504.

[26] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, Mar. 2004, https://doi.org/10.1145/1007730.1007733.

[27] C. Kim and A. Ferrara, Eds., *Gestational Diabetes During and After Pregnancy*. London, UK: Springer, 2010.

[28] S. M. Camhi *et al.*, "The Relationship of Waist Circumference and BMI to Visceral, Subcutaneous, and Total Body Fat: Sex and Race Differences," *Obesity*, vol. 19, no. 2, pp. 402–408, 2011, https://doi.org/10.1038/oby.2010.248.

[29] S. E. Kahn, R. L. Hull, and K. M. Utzschneider, "Mechanisms linking obesity to insulin resistance and type 2 diabetes," *Nature*, vol. 444, no. 7121, pp. 840–846, Dec. 2006, https://doi.org/10.1038/nature05482.

[30] "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021," *Diabetes Care*, vol. 44, no. s1, pp. S15–S33, Dec. 2020, https://doi.org/10.2337/dc21-S002.

[31] J. R. Sowers, M. Epstein, and E. D. Frohlich, "Diabetes, Hypertension, and Cardiovascular Disease," *Hypertension*, vol. 37, no. 4, pp. 1053–1059, Apr. 2001, https://doi.org/10.1161/01.HYP.37.4.1053.

[32] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[33] S. van Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45, pp. 1–67, Dec. 2011, https://doi.org/10.18637/jss.v045.i03.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, https://doi.org/10.1613/jair.953.

[35] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging Artificial Intelligence Applications in Computer Engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[36] T. Widiyaningtyas, H. Hairani, D. D. Prasetya, U. Pujianto, and W. Caesarendra, "A Modified SMOTE with Noise Filtering and Manhattan Distance Metric Approach to Address Imbalanced Health Datasets," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25452–25459, Aug. 2025, https://doi.org/10.48084/etasr.11925.

[37] M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi, "Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset," *Fuzzy Information and Engineering*, vol. 9, no. 3, pp. 345–357, Sep. 2017, https://doi.org/10.1016/j.fiae.2017.09.006.

[38] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, Dec. 1997, https://doi.org/10.1016/S0004-3702(97)00043-X.

[39] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, May 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.