PREMIER UNIVERSITY, CHATTOGRAM

Department of Computer Science & Engineering

Project Report

Machine Learning for Diabetes Prediction on DiaBD Dataset

Submitted by

Name: Sanjida Mahmud Muntaha
ID: 2104010202279

Name: Tainur Rahaman
ID: 2104010202291

Name: Akibul Islam
ID: 2104010202299

Section: D   Batch: 40th

Submitted to

MD Tamim Hossain
Lecturer
Department of Computer Science & Engineering
Premier University, Chattogram

November 23, 2025

Contents

Machine Learning for Diabetes Prediction on DiaBD Dataset

Abstract

Diabetes mellitus imposes a significant global health burden, making early detection strategies vital. This study evaluates machine learning models for diabetes prediction using the DiaBD dataset from Bangladesh, comprising 5,288 clinical records. We implemented a rigorous preprocessing pipeline involving BMI integrity verification, winsorization, and Isolation Forest for multivariate outlier detection. Severe class imbalance (14.46:1) was addressed through class weighting strategies to preserve data integrity. Feature engineering yielded 18 clinically relevant variables, including pulse pressure and log-transformed glucose.

We evaluated Random Forest, XGBoost, and Logistic Regression using 5-fold stratified cross-validation. Random Forest achieved the optimal performance with an AUC-ROC of 0.8463, 86.39% accuracy, and 61.76% recall. While XGBoost demonstrated superior sensitivity (69.12% recall), Random Forest provided the most balanced metrics. Despite low precision (∼26%) resulting from the extreme imbalance, this work demonstrates that carefully engineered features and appropriate weighting techniques can produce clinically viable prediction models.

Keywords: Diabetes Prediction, Machine Learning, Class Imbalance, Feature Engineering, Random Forest, DiaBD Dataset

1 Introduction and Problem Statement

Diabetes mellitus is a chronic metabolic disorder affecting over 537 million adults globally as of 2021, with projections indicating a rise to 783 million by 2045 [1]. The disease imposes substantial healthcare burdens, particularly in low- and middle-income countries where diagnostic infrastructure remains limited. Bangladesh, with a population exceeding 170 million, faces a rapidly growing diabetes epidemic, with prevalence rates increasing from 5% in 2000 to approximately 10% in recent years [2]. Early detection is critical for preventing severe complications including cardiovascular disease, kidney failure, and neuropathy, yet access to specialized diagnostic services remains constrained in resource-limited settings.

Machine learning offers a promising avenue for scalable, cost-effective diabetes risk assessment by leveraging routinely collected clinical data. However, developing robust predictive models for medical applications presents significant challenges, including severe class imbalance, data quality issues, and the need for clinical interpretability. This study addresses these challenges through a comprehensive machine learning pipeline applied to the DiaBD dataset [3], a collection of 5,288 clinical records from Bangladesh encompassing demographic, anthropometric, vital sign, laboratory, and medical history features.

1.1 Problem Statement

The primary objective of this research is to develop and evaluate machine learning models capable of accurately predicting diabetes status from clinical measurements while addressing the following challenges:

1. **Extreme Class Imbalance:** The dataset exhibits a 14.46:1 ratio of non-diabetic to diabetic cases (93.53% vs. 6.47%), requiring specialized techniques to prevent model bias toward the majority class.
2. **Data Quality and Integrity:** Medical datasets frequently contain measurement errors, inconsistencies (e.g., miscalculated BMI values), and physiologically impossible values that must be identified and corrected to ensure model reliability.
3. **Clinical Validity:** Feature engineering must incorporate domain knowledge to create medically meaningful predictors (e.g., pulse pressure, hypertension thresholds) rather than relying solely on raw measurements.
4. **Model Interpretability:** Healthcare applications demand transparent models where feature importance can be understood and validated by medical practitioners.

2 Related Work

Machine learning approaches for diabetes prediction have evolved, focusing on leveraging routine clinical data. Early studies utilized classical algorithms like **Random Forest** and **Support Vector Machines (SVM)** on datasets like the Pima Indians Diabetes Database (PIDD), achieving moderate accuracy but often overlooking class imbalance [4]. **Logistic Regression** remains valuable due to its high interpretability in clinical settings [5].

Recent research favors **ensemble boosting methods** for their high performance on structured medical data. **XGBoost** has demonstrated superior predictive capability in large cohort studies, achieving AUC-ROC scores up to 0.86 through careful hyperparameter tuning and domain-specific feature incorporation [6].

A fundamental challenge in this field is **extreme class imbalance**, as diabetes prevalence is low. Researchers employ two primary strategies:

- **Synthetic Oversampling (SMOTE):** Generates artificial data points for the minority class, though this risks introducing unrepresentative patterns [7].
- **Cost-Sensitive Learning:** Assigns higher misclassification penalties to the diabetic (minority) class. This approach, often implemented via class weighting in algorithms like Random Forest, has successfully improved diabetes **recall** (sensitivity) without compromising data authenticity [8].

The effectiveness of these models hinges on **Feature Engineering**, which transforms raw measurements into clinically meaningful predictors (e.g., **pulse pressure, BMI categories**). Integrating domain knowledge this way significantly enhances model performance and interpretability [9].

2.1 Research Gap

While progress is evident, several gaps persist: (1) inadequate focus on **data quality and integrity** specific to medical datasets, (2) limited comparison of imbalance handling strategies that prioritize data authenticity (i.e., class weighting vs. oversampling), and (3) sparse evaluation of predictive models on **South Asian populations** where the diabetes burden is accelerating rapidly. This study addresses these gaps through a rigorous methodology applied to the DiaBD dataset from Bangladesh [3].

3 Dataset

3.1 Data Source and Description

This study utilizes the DiaBD (Diabetes Bangladesh) dataset [3], publicly available on Mendeley Data. The dataset comprises 5,288 clinical records collected from healthcare facilities in Bangladesh, representing a diverse population sample with varying diabetes risk profiles. Each record contains 14 features spanning demographic, anthropometric, vital sign, laboratory, and medical history domains, along with a binary diabetes diagnosis label.

Table 1 summarizes the dataset characteristics. Notably, the dataset exhibits no missing values and no duplicate records, indicating high-quality data collection practices. However, as detailed in subsequent sections, significant data integrity issues were identified that required systematic correction.

Table 1 Dataset Overview and Characteristics

Characteristic	**Value**
Total samples	5,288
Features (excluding target)	14
Target variable	Diabetic (Yes/No)
Missing values	0 (0%)
Duplicate records	0 (0%)
Diabetic cases	342 (6.47%)
Non-diabetic cases	4,946 (93.53%)
Class imbalance ratio	14.46:1

The 14 original features are categorized as follows:

- **Demographic (2):** Age (years), Gender (Male/Female)
- **Anthropometric (3):** Height (meters), Weight (kg), BMI (kg/m^2)
- **Vital Signs (3):** Pulse rate (bpm), Systolic blood pressure (mmHg), Diastolic blood pressure (mmHg)

- **Laboratory (1):** Glucose (mmol/L)
- **Medical History (5):** Family diabetes, Hypertensive status, Family hypertension, Cardiovascular disease, Stroke history

3.2 Class Distribution and Imbalance

Figure 1 illustrates the severe class imbalance inherent in the dataset. With only 6.47% positive diabetes cases, the dataset reflects realistic population-level prevalence but poses significant challenges for machine learning model training. This 14.46:1 imbalance ratio necessitates specialized techniques to prevent models from trivially predicting the majority class.

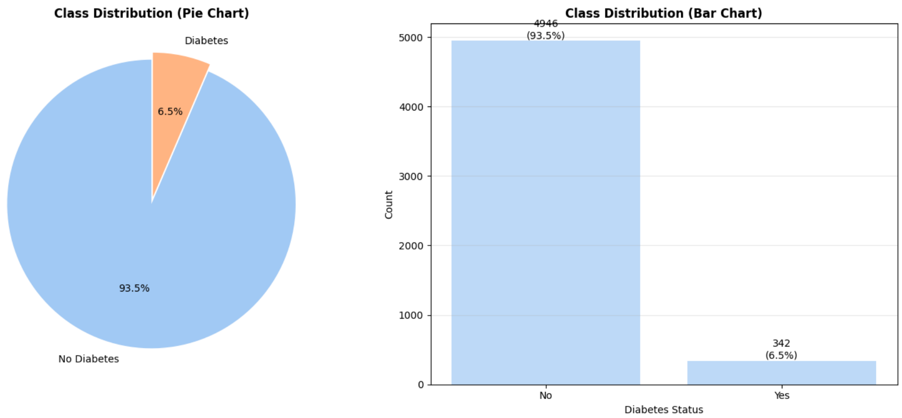


Fig. 1 Target variable distribution showing extreme class imbalance (14.46:1 ratio)

3.3 Exploratory Data Analysis

3.3.1 Descriptive Statistics

Table 2 presents descriptive statistics for all numerical features. Several observations warrant attention:

- **Physiologically impossible values:** Pulse rate minimum of 5 bpm, height of 0.36 m, weight of 3 kg, and BMI of 560.19 indicate data entry errors requiring systematic filtering.
- **Extreme outliers:** The maximum BMI (560.19) versus the 75th percentile (24.49) reveals catastrophic outliers, likely resulting from incorrect unit conversions or typographical errors.
- **Zero glucose:** Glucose values of 0.00 mmol/L are biologically incompatible with life, confirming the presence of invalid measurements.

Figure 2 visualizes the distribution of original features through boxplots, clearly revealing the extent of outlier contamination across multiple variables.

Table 2 Descriptive Statistics of Numerical Features

Feature	Mean	Std	Min	25%	Median	75%	Max
Age	45.75	13.42	21.00	35.00	45.00	55.00	80.00
Pulse Rate	76.63	12.23	5.00	69.00	76.00	84.00	133.00
Systolic BP	134.00	22.23	62.00	119.00	130.00	147.00	231.00
Diastolic BP	82.23	12.48	45.00	73.00	81.00	90.00	119.00
Glucose	7.56	2.94	0.00	6.00	6.93	8.13	33.46
Height	1.55	0.08	0.36	1.52	1.55	1.60	1.96
Weight	53.64	10.08	3.00	46.77	53.00	59.90	100.70
BMI	22.50	8.65	1.22	19.64	21.91	24.49	560.19

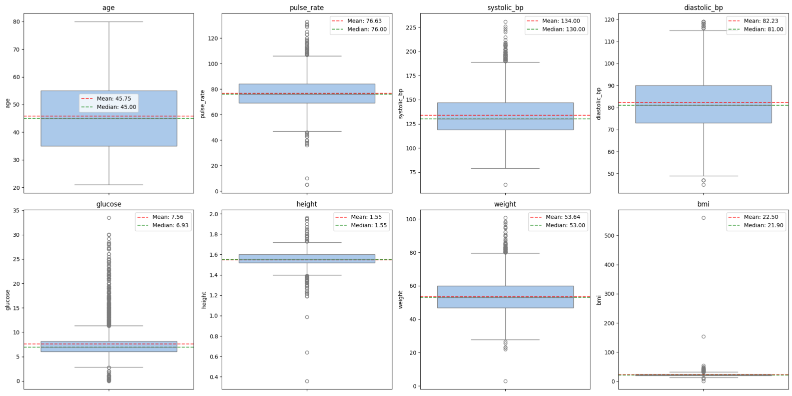


Fig. 2 Boxplots of original features showing extensive outliers and distribution skewness

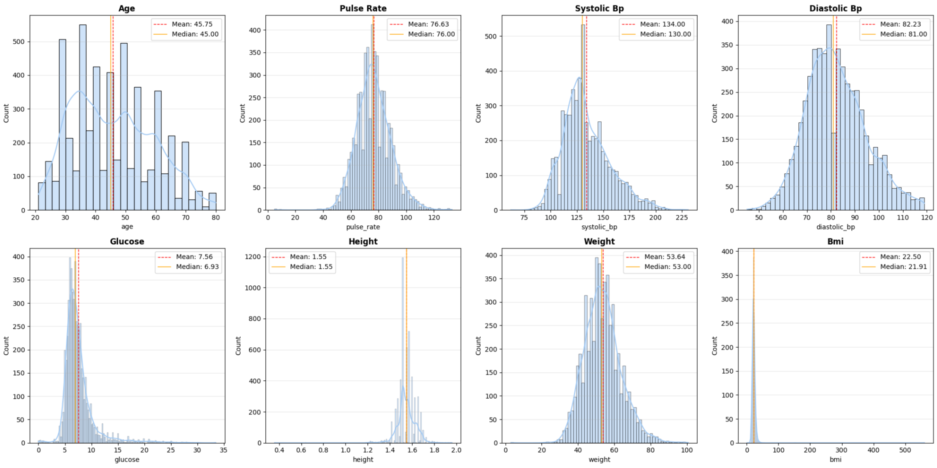


Fig. 3 Distributions of continuous clinical features with mean (dashed) and median (solid) lines.

3.3.2 Feature Correlations

Table 3 lists features ranked by their absolute correlation with the diabetes target variable. Glucose demonstrates the strongest predictive relationship (r = 0.306), followed by blood pressure measurements. Notably, BMI exhibits relatively weak direct correlation (r = 0.044), suggesting that raw BMI alone may be insufficient without categorical transformations.

Table 3 Features Ranked by Correlation with Diabetes Status

Feature	Absolute Correlation
Glucose	0.3056
Diastolic BP	0.1559
Systolic BP	0.1559
Weight	0.1079
Age	0.0902
Pulse Rate	0.0507
BMI	0.0437
Height	0.0309

3.4 Data Preprocessing Pipeline

A rigorous multi-stage preprocessing pipeline was implemented to address data quality issues while preventing information leakage. All preprocessing steps were applied exclusively to the training set, with learned parameters subsequently applied to the test set.

3.4.1 Train-Test Split

An 80-20 stratified split was performed, yielding 4,230 training samples and 1,058 test samples while preserving the 14.46:1 class distribution in both sets. This split was executed *before* any preprocessing to ensure test set integrity.

3.4.2 BMI Integrity Verification

Critical to medical data validation, we verified that reported BMI values matched calculated values using the formula BMI = $\frac{\text{weight (kg)}}{\text{height (m)}^2}$. A tolerance threshold of 0.1 kg/m^2 identified 1,954 records (36.95%) with erroneous BMI values. Maximum deviation reached 13.94 kg/m^2, confirming systematic data entry errors. All incorrect BMI values were replaced with properly calculated values before subsequent processing.

3.4.3 Medical Validity Filters

Physiologically impossible values were removed from the training set using clinically informed thresholds (Table 4). These filters eliminated 83 training samples (1.96%), reducing the training set from 4,230 to 4,147 samples. The test set remained unfiltered to preserve real-world deployment conditions.

Table 4 Medical Validity Filter Criteria

Feature	Valid Range
Height	1.0–2.2 meters
Weight	30–200 kg
Pulse Rate	40–180 bpm
BMI	10–60 kg/m^2
Glucose	2–20 mmol/L

3.4.4 Winsorization

To mitigate remaining outliers without data loss, winsorization was applied to seven features (systolic BP, diastolic BP, pulse rate, glucose, BMI, height, weight) by clipping values at the 1st and 99th percentiles calculated from the training set. Table 5 shows the computed bounds, which were applied identically to both training and test sets.

Table 5 Winsorization Bounds (1st–99th Percentile)

Feature	1st Percentile	99th Percentile
Systolic BP	95.00	196.00
Diastolic BP	56.00	115.54
Pulse Rate	51.00	109.00
Glucose	4.04	17.06
BMI	14.79	34.40
Height	1.35	1.73
Weight	33.70	80.91

3.4.5 Clinical Feature Engineering

Domain knowledge was incorporated through four engineered features:

1. **Pulse Pressure:** systolic_bp − diastolic_bp, a cardiovascular risk indicator strongly associated with arterial stiffness and diabetes complications.
2. **Hypertension Flag:** Binary indicator (1 if systolic BP $\geq$ 140 mmHg OR diastolic BP $\geq$ 90 mmHg), based on clinical hypertension diagnostic criteria.
3. **Glucose Log:** $\log(1 + \text{glucose})$ transformation to reduce right-skewness and stabilize variance.
4. **BMI Categories:** WHO-standard classifications (Underweight: <18.5, Normal: 18.5–24.9, Overweight: 25–29.9, Obese: $\geq$30), one-hot encoded into four binary features.

3.4.6 Multicollinearity Resolution

To prevent redundancy, three features were removed: (1) original glucose (replaced by glucose_log), (2) height and weight (captured by BMI), and (3) obesity_flag (perfectly correlated with bmi_Obese dummy variable). Gender was encoded as Male=1, Female=0.

3.4.7 Multivariate Outlier Detection

Isolation Forest was applied to five features (age, pulse_rate, bmi, glucose_log, pulse_pressure) with contamination parameter 0.05, detecting 208 multivariate outliers (5.0% of 4,147 training samples). These samples were removed, yielding a final training set of 3,939 samples with 18 features. The test set remained unmodified at 1,058 samples.

Table 6 summarizes the complete preprocessing pipeline impact:

Table 6 Preprocessing Pipeline Summary

Step	Input Size	Output Size	Removed
Train-Test Split (80-20)	5,288	4,230 (train)	1,058 (test)
Medical Validity Filters	4,230	4,147	83
Isolation Forest (5%)	4,147	3,939	208
Final Training Set	–	**3,939**	**291 total**
Test Set (unchanged)	–	**1,058**	**0**

After preprocessing, the training set imbalance increased to 18.03:1 (3,732 non-diabetic vs. 207 diabetic), necessitating robust imbalance handling strategies detailed in Section 5.

3.5 Final Feature Set

After the complete preprocessing pipeline, the final dataset comprises 18 features organized into three categories based on their data types and roles in the predictive model.

3.5.1 Continuous Numerical Features

These features represent measurements on continuous scales and capture the core physiological and demographic characteristics:

- **age**: Patient age in years
- **pulse_rate**: Heart rate in beats per minute (winsorized)
- **systolic_bp**: Systolic blood pressure in mmHg (winsorized)
- **diastolic_bp**: Diastolic blood pressure in mmHg (winsorized)
- **bmi**: Body Mass Index in kg/m^2 (corrected and winsorized)
- **pulse_pressure**: Engineered feature (systolic_bp − diastolic_bp)
- **glucose_log**: Log-transformed glucose, $\log(1 + \text{glucose})$

3.5.2 Binary Categorical Features

These features represent binary medical conditions or demographic characteristics:

- **gender**: Male (1) or Female (0)
- **family_diabetes**: Family history of diabetes
- **hypertensive**: Current hypertensive diagnosis
- **family_hypertension**: Family history of hypertension
- **cardiovascular_disease**: History of cardiovascular disease
- **stroke**: History of stroke
- **hypertension_flag**: Engineered feature indicating BP $\geq$ 140/90 mmHg

3.5.3 One-Hot Encoded Features

BMI categories based on WHO classifications, one-hot encoded:

- **bmi_Underweight**: BMI $<$ 18.5 kg/m^2
- **bmi_Normal**: BMI 18.5–24.9 kg/m^2
- **bmi_Overweight**: BMI 25–29.9 kg/m^2
- **bmi_Obese**: BMI $\geq$ 30 kg/m^2

Note that `obesity_flag` (originally engineered) was removed during correlation cleanup as it exhibited perfect multicollinearity (r = 1.0) with `bmi_Obese`.

3.5.4 Post-Preprocessing Feature Distributions

Figures 4 and 5 illustrate the distributions of continuous numerical features after all preprocessing steps (medical filtering, winsorization, and outlier removal). These visualizations confirm successful outlier mitigation while preserving clinically meaningful variance. The features exhibit reduced skewness compared to raw measurements, with glucose_log demonstrating notably improved distributional properties relative to the original right-skewed glucose variable.

Table 7 presents the final feature-target correlations after all preprocessing and engineering steps. Notably, the `hypertensive` binary feature exhibits the strongest correlation (r = 0.354), while `glucose_log` (r = 0.214) demonstrates improved predictive signal compared to the original glucose variable. The engineered `pulse_pressure` and `hypertension_flag` features show moderate correlations, validating their clinical relevance.

The preprocessing pipeline successfully transformed the raw dataset into a clean, feature-rich representation suitable for machine learning model training while preserving test set integrity for unbiased performance evaluation.

3.6 Multicollinearity Detection and Resolution

Feature correlation analysis was performed in two stages to identify and resolve multicollinearity issues that could degrade model performance and interpretability.

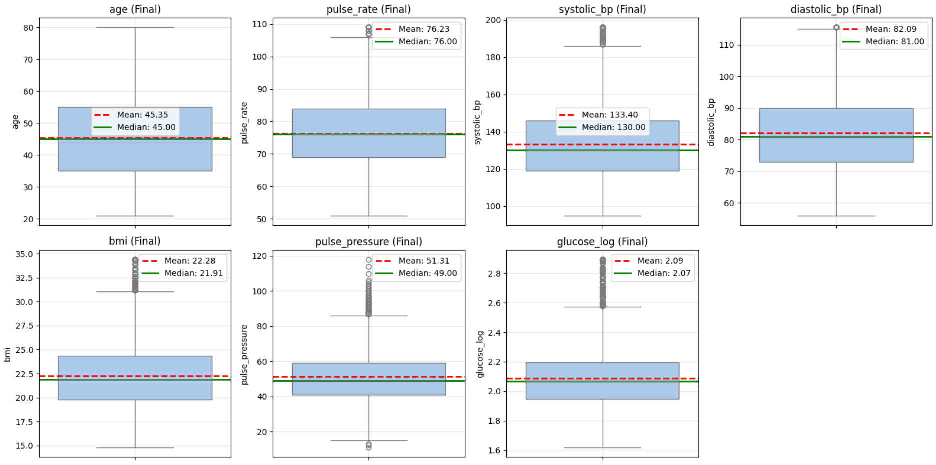


Fig. 4 Boxplots of final numerical features after preprocessing, showing mean (red dashed) and median (green solid) lines. Outliers have been effectively controlled through winsorization and multivariate detection.

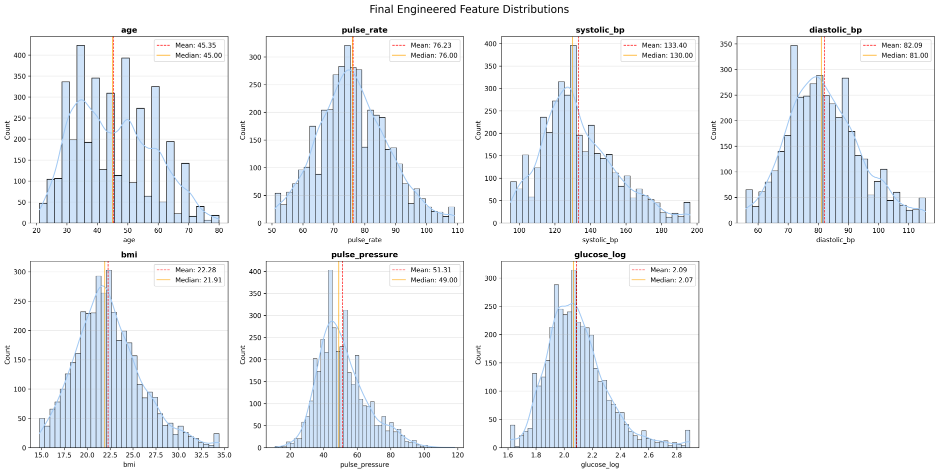


Fig. 5 Histograms with kernel density estimation (KDE) for final numerical features. Mean (red dashed) and median (orange solid) lines highlight central tendencies. The glucose_log transformation successfully reduced right-skewness present in the original glucose distribution.

3.6.1 Stage 1: Initial Correlation Analysis

After feature engineering but before cleanup, the correlation matrix revealed two critical multicollinearity concerns (Table 8):

Perfect Multicollinearity Detected: The engineered `obesity_flag` (defined as BMI ≥ 30) exhibited perfect correlation ($r = 1.0$) with the one-hot encoded `bmi_Obese` category, representing complete redundancy. This arose because both features encoded identical information—the WHO obesity threshold—through different representations (binary flag vs. categorical dummy variable).

Table 7 Final Features Ranked by Correlation with Diabetes Status

Feature	Correlation
hypertensive	0.354
glucose_log	0.214
systolic_bp	0.162
diastolic_bp	0.159
hypertension_flag	0.148
bmi	0.100
pulse_pressure	0.100
age	0.097
cardiovascular_disease	0.078
bmi_Overweight	0.077
bmi_Underweight	-0.064

Table 8 High Feature Correlations Before Cleanup (|r| > 0.85)

Feature 1	Feature 2	Correlation
family_diabetes	family_hypertension	0.9579
obesity_flag	bmi_Obese	1.0000

Resolution Strategy: To eliminate this redundancy, `obesity_flag` was removed from both training and test sets, retaining the more granular BMI categorical encoding (`bmi_Underweight`, `bmi_Normal`, `bmi_Overweight`, `bmi_Obese`) which captures all weight classifications rather than a single threshold.

3.6.2 Stage 2: Post-Cleanup Correlation Analysis

After removing `obesity_flag`, the final 18-feature dataset was re-examined for remaining multicollinearity issues. Table 9 summarizes the results.

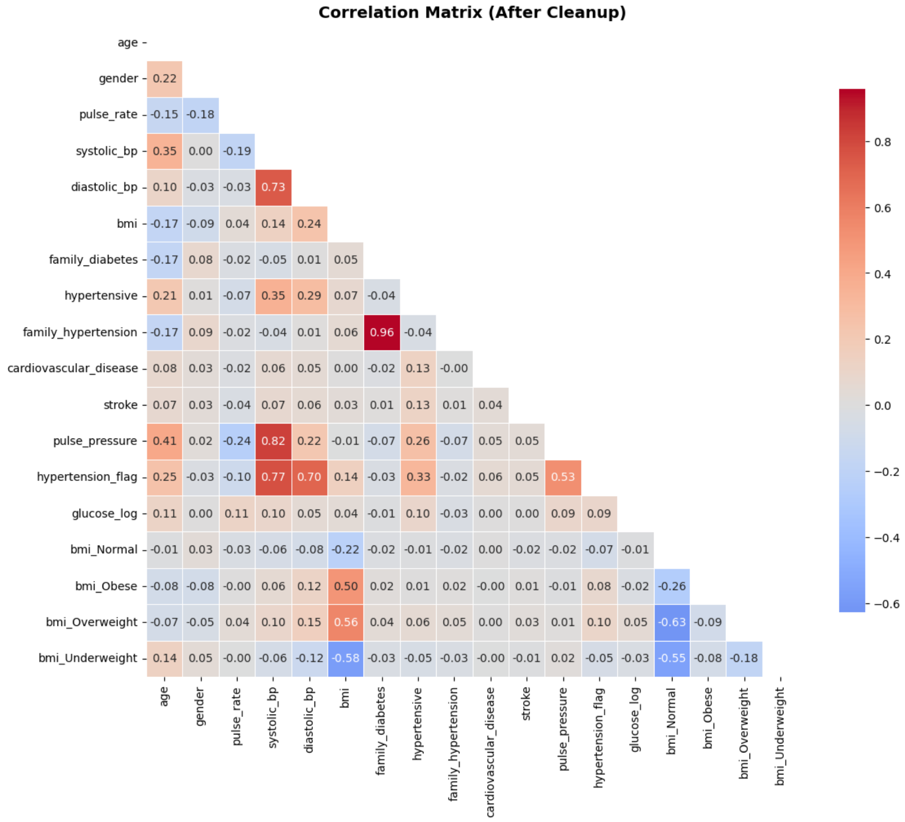
Table 9 High Feature Correlations After Cleanup (|r| > 0.85)

Feature 1	Feature 2	Correlation
family_diabetes	family_hypertension	0.9579

Only one high correlation remains: `family_diabetes` and `family_hypertension` (r = 0.958). This strong association reflects genuine clinical comorbidity patterns rather than artificial redundancy. Families sharing genetic predisposition and environmental factors (diet, physical activity, socioeconomic status) commonly exhibit clustering of both conditions. Importantly, these represent distinct medical diagnoses with independent clinical significance:

- **Retained both features**: Unlike the obesity_flag case, these features capture different aspects of family medical history despite high correlation.
- **Clinical justification**: Family diabetes history specifically predicts genetic susceptibility to impaired glucose metabolism, while family hypertension history indicates inherited cardiovascular risk factors.
- **Algorithmic resilience**: Tree-based ensemble methods (Random Forest, XGBoost) employed in this study are inherently robust to correlated features, as they can partition data using either feature and capture interaction effects.

Figure 6 presents the final correlation matrix, confirming that all other feature pairs exhibit correlations well below the 0.85 threshold, indicating successful multicollinearity control.



	age	gender	pulse_rate	systolic_bp	diastolic_bp	bmi	family_diabetes	hypertensive	family_hypertension	cardiovascular_disease	stroke	pulse_pressure	hypertension_flag	glucose_log	bmi_Normal	bmi_Obese	bmi_Overweight	bmi_Underweight
age																		
gender	0.22																	
pulse_rate	-0.15	-0.18																
systolic_bp	0.35	0.00	-0.19															
diastolic_bp	0.10	-0.03	-0.03	0.73														
bmi	-0.17	-0.09	0.04	0.14	0.24													
family_diabetes	-0.17	0.08	-0.02	-0.05	0.01	0.05												
hypertensive	0.21	0.01	-0.07	0.35	0.29	0.07	-0.04											
family_hypertension	-0.17	0.09	-0.02	-0.04	0.01	0.06	0.96	-0.04										
cardiovascular_disease	0.08	0.03	-0.02	0.06	0.05	0.00	-0.02	0.13	-0.00									
stroke	0.07	0.03	-0.04	0.07	0.06	0.03	0.01	0.13	0.01	0.04								
pulse_pressure	0.41	0.02	-0.24	0.82	0.22	-0.01	-0.07	0.26	-0.07	0.05	0.05							
hypertension_flag	0.25	-0.03	-0.10	0.77	0.70	0.14	-0.03	0.33	-0.02	0.06	0.05	0.53						
glucose_log	0.11	0.00	0.11	0.10	0.05	0.04	-0.01	0.10	-0.03	0.00	0.00	0.09	0.09					
bmi_Normal	-0.01	0.03	-0.03	-0.06	-0.08	-0.22	-0.02	-0.01	-0.02	0.00	-0.02	-0.02	-0.07	-0.01				
bmi_Obese	-0.08	-0.08	-0.00	0.06	0.12	0.50	0.02	0.01	0.02	-0.00	0.01	-0.01	0.08	-0.02	-0.26			
bmi_Overweight	-0.07	-0.05	0.04	0.10	0.15	0.56	0.04	0.06	0.05	0.00	0.03	0.01	0.10	0.05	-0.63	-0.09		
bmi_Underweight	0.14	0.05	-0.00	-0.06	-0.12	-0.58	-0.03	-0.05	-0.03	-0.00	-0.01	0.02	-0.05	-0.03	-0.55	-0.08	-0.18	

Fig. 6 Final correlation matrix after removing obesity_flag. Only one high correlation remains (family diabetes indicators at r = 0.958), which was retained for clinical validity.

3.6.3 Feature-Target Relationship Analysis

Figure 7 visualizes the top 10 features ranked by absolute correlation with diabetes status in the final cleaned dataset. The `hypertensive` binary feature demonstrates the strongest linear relationship (r = 0.354), consistent with extensive epidemiological evidence linking hypertension and type 2 diabetes through shared pathophysiological mechanisms including insulin resistance and endothelial dysfunction.

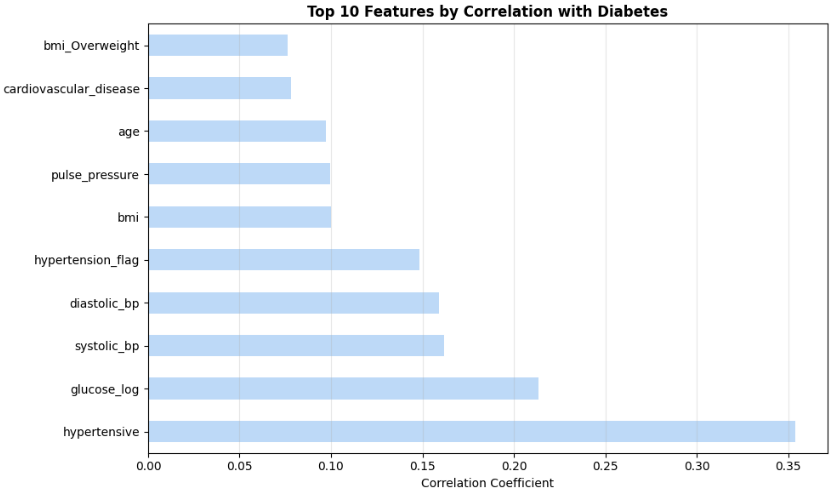


Fig. 7 Top 10 features by absolute correlation with diabetes outcome. Positive correlations (red) indicate increased diabetes risk, while negative correlations (blue) suggest protective associations.

Key observations from the final correlation analysis:

- **Glucose_log leads biochemical markers** (r = 0.214): The log transformation preserved glucose's predictive signal while improving distributional properties for linear models.
- **Blood pressure cluster**: `systolic_bp` (r = 0.162), `diastolic_bp` (r = 0.159), and engineered `hypertension_flag` (r = 0.148) form a coherent set of cardiovascular risk indicators.
- **BMI and pulse_pressure equivalence** (both r = 0.100): Despite weak direct correlation, BMI categories may capture nonlinear obesity-diabetes relationships not evident in linear correlation.
- **Age shows moderate correlation** (r = 0.097): Consistent with diabetes prevalence increasing through progressive beta-cell dysfunction and accumulated metabolic risk.
- **BMI_Underweight exhibits negative correlation** (r = -0.064): This protective association reflects lower metabolic syndrome prevalence, though interpretation requires caution due to potential confounding by malnutrition or other health conditions.

Notably, no single feature exhibits extremely high correlation (r > 0.5) with the target, indicating that accurate diabetes prediction requires integrating multiple complementary signals—a task well-suited to the ensemble machine learning methods evaluated in this study. The preprocessing pipeline successfully transformed the raw dataset into a clean, 18-feature representation with controlled multicollinearity, ready for model training.

4 Methodology

This section details the machine learning algorithms, hyperparameter optimization strategy, and class imbalance handling techniques employed to develop robust diabetes prediction models.

4.1 Algorithm Selection

Three classification algorithms representing distinct modeling paradigms were selected to provide comprehensive performance comparison:

4.1.1 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of individual tree predictions. Key advantages for this medical application include:

- **Robustness to outliers**: Despite preprocessing, residual outliers in clinical data are inherently tolerated through bootstrap aggregation.
- **Feature importance quantification**: Provides interpretable rankings of feature contributions via mean decrease in impurity.
- **Nonlinear relationship capture**: Recursive partitioning naturally models complex interactions without explicit feature engineering.
- **Reduced overfitting**: Ensemble averaging mitigates individual tree variance, particularly valuable given the modest training set size (3,939 samples).

4.1.2 XGBoost

XGBoost (Extreme Gradient Boosting) implements gradient boosting with advanced regularization techniques. This algorithm was selected for:

- **Superior predictive performance**: Consistently achieves state-of-the-art results on structured/tabular data across diverse domains.
- **Built-in imbalance handling**: The `scale_pos_weight` parameter directly addresses class imbalance without synthetic data generation.
- **Regularization capabilities**: L1 (lasso) and L2 (ridge) penalties, combined with tree pruning, prevent overfitting on small datasets.
- **Computational efficiency**: Histogram-based tree construction and GPU acceleration enable rapid hyperparameter exploration.

4.1.3 Logistic Regression

Logistic Regression serves as both a competitive baseline and interpretable linear model. Its inclusion provides:

- **Clinical interpretability**: Coefficients directly quantify feature-target relationships as log-odds ratios, facilitating medical validation.
- **Computational simplicity**: Rapid training enables extensive hyperparameter search despite the 30-iteration budget.
- **Probabilistic outputs**: Naturally produces calibrated probability estimates critical for risk stratification in clinical decision support.
- **Regularization options**: L1 and L2 penalties enable feature selection and multicollinearity mitigation.

4.2 Class Imbalance Handling

The extreme class imbalance (18.03:1 ratio after preprocessing) necessitated careful strategy selection. Rather than employing synthetic oversampling techniques (e.g., SMOTE), which risk introducing artificial patterns not representative of true diabetes presentations, we adopted **class weighting**—a mathematically principled approach that adjusts the loss function during training.

4.2.1 Class Weight Calculation

Balanced class weights were computed using scikit-learn's `compute_class_weight` function with the `'balanced'` strategy:

$$w_i = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_{\text{samples in class } i}} \tag{1}$$

For our training set:

- w_0 (No Diabetes) = 0.528
- w_1 (Diabetes) = 9.514

This 18-fold weighting amplifies the gradient contribution of diabetes samples, effectively signaling to models that minority class misclassifications incur substantially higher penalties.

4.2.2 Algorithm-Specific Implementation

- **Random Forest & Logistic Regression**: Class weights passed via the `class_weight='balanced'` parameter, which internally applies weights to sample contributions in tree splitting criteria (Random Forest) or loss function (Logistic Regression).
- **XGBoost**: Implemented through `scale_pos_weight=18.03`, a specialized parameter that adjusts the gradient contribution of positive class samples during boosting iterations. This differs from standard class weighting but achieves equivalent effect for binary classification.

4.3 Feature Scaling

All features were standardized using Z-score normalization to ensure fair contribution across different measurement scales:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

where μ and σ denote the mean and standard deviation computed exclusively from the training set. The fitted `StandardScaler` was applied identically to both training and test sets, preventing data leakage. While tree-based methods (Random Forest, XGBoost) are scale-invariant, standardization benefits Logistic Regression and facilitates cross-algorithm comparison.

4.4 Hyperparameter Optimization

A randomized search strategy with 30 iterations per algorithm was employed, balancing exploration breadth with computational efficiency. This approach randomly samples from specified distributions, empirically demonstrating superior performance-to-cost ratio compared to grid search.

4.4.1 Cross-Validation Protocol

5-fold stratified cross-validation ensured:

- **Stratification**: Each fold preserved the 18.03:1 class ratio, preventing degenerate folds with zero minority class samples.
- **Unbiased evaluation**: ROC-AUC scores averaged across folds provided robust estimates of generalization performance.
- **Reproducibility**: Fixed random seed (42) ensures deterministic fold splits across runs.

4.4.2 Optimization Objective

ROC-AUC (Area Under the Receiver Operating Characteristic Curve) served as the primary optimization metric due to its:

- **Threshold independence**: Evaluates model performance across all classification thresholds, crucial for medical applications where operating points vary by clinical context.
- **Imbalance robustness**: Unlike accuracy, ROC-AUC remains informative under severe class imbalance by separately quantifying true positive and false positive rates.
- **Probabilistic interpretation**: Represents the probability that a randomly chosen diabetic patient receives a higher prediction score than a randomly chosen non-diabetic patient.

4.4.3 Hyperparameter Search Spaces

Tables 10, 11, and 12 detail the hyperparameter search spaces for each algorithm.

Rationale for Search Ranges:

Table 10 Random Forest Hyperparameter Search Space

Parameter	**Search Space**
n_estimators	randint(100, 500)
max_depth	randint(5, 21)
min_samples_split	randint(2, 21)
min_samples_leaf	randint(1, 11)
max_features	{sqrt, log2, None}
Fixed parameters:	
class_weight	'balanced'
random_state	42
n_jobs	-1 (all cores)

Table 11 Logistic Regression Hyperparameter Search Space

Parameter	**Search Space**
C	uniform(0.001, 10.001)
penalty	{L1, L2}
Fixed parameters:	
solver	'liblinear'
class_weight	'balanced'
max_iter	1000
random_state	42

- **n_estimators (RF: 100–500, XGB: 100–300)**: Balances ensemble diversity with computational cost. Higher values increase robustness but offer diminishing returns beyond 300–500 trees.
- **max_depth (RF: 5–21, XGB: 3–10)**: Controls model complexity. Shallow trees prevent overfitting on small datasets; deeper trees capture intricate interactions.
- **learning_rate (XGB: 0.01–0.31)**: Governs boosting step size. Lower values require more trees but improve generalization through gradual learning.
- **Regularization parameters**: `min_samples_split/leaf` (RF), `min_child_weight/gamma` (XGB), and `C/penalty` (LR) collectively control model complexity, with ranges calibrated to dataset size.

4.5 Model Selection and Finalization

For each algorithm, the hyperparameter configuration achieving the highest mean cross-validation ROC-AUC across 30 randomized trials was selected. The winning configuration was then retrained on the entire training set (3,939 samples) to produce the final model. This two-stage process (optimization → full retraining) maximizes utilization of available training data while maintaining unbiased hyperparameter selection.

Table 12 XGBoost Hyperparameter Search Space

Parameter	Search Space
n_estimators	randint(100, 301)
max_depth	randint(3, 11)
learning_rate	uniform(0.01, 0.31)
subsample	uniform(0.6, 1.0)
colsample_bytree	uniform(0.6, 1.0)
min_child_weight	randint(1, 11)
gamma	uniform(0, 0.5)
tree_method	'hist'
Fixed parameters:	
scale_pos_weight	18.03
eval_metric	'logloss'
device	'cuda' (if available)
random_state	42

4.6 Implementation Details

- **Software**: Python 3.10 with scikit-learn 1.3.0, XGBoost 2.0.3, pandas 2.0.3, NumPy 1.24.3
- **Hardware**: Google Colab environment with NVIDIA T4 GPU (16GB VRAM) for XGBoost acceleration
- **Reproducibility**: All random operations seeded with `RANDOM_SEED=42` for deterministic results
- **Parallelization**: Random Forest and cross-validation leveraged multi-core CPU (`n_jobs=-1`) for 4–8× speedup

The complete methodology ensures that model development adheres to best practices for medical machine learning: rigorous preprocessing, principled imbalance handling, unbiased hyperparameter optimization, and strict train-test separation to produce reliable generalization estimates.

5 Training Procedure

This section documents the complete training workflow, including optimization results, computational environment, and reproducibility considerations.

5.1 Hyperparameter Optimization Results

Randomized search with 30 iterations per algorithm was executed using 5-fold stratified cross-validation on the training set (3,939 samples, 18 features). Table 13 summarizes the best cross-validation ROC-AUC scores achieved for each algorithm.

Notably, cross-validation scores (0.77–0.79) were systematically lower than final test set performance (0.81–0.85, detailed in Section 7), indicating that models generalized *better* than expected—a favorable outcome suggesting conservative hyperparameter selection and absence of overfitting.

Table 13 Cross-Validation Optimization Results

Algorithm	Best CV ROC-AUC	Iterations
XGBoost	0.7900	30
Logistic Regression	0.7865	30
Random Forest	0.7706	30

5.1.1 XGBoost Optimal Configuration

XGBoost achieved the highest cross-validation ROC-AUC (0.7900) among all algorithms. The optimized configuration (Table 14) reflects a conservative parameterization: shallow trees (depth=3), low learning rate (0.038), and moderate ensemble size (101 estimators) prioritize generalization over training set fit.

Table 14 XGBoost: Best Hyperparameters from Cross-Validation

Parameter	Optimized Value
n_estimators	101
max_depth	3
learning_rate	0.0379
subsample	0.9589
colsample_bytree	0.9602
min_child_weight	7
gamma	0.1695
tree_method	'hist'
Fixed parameters:	
scale_pos_weight	18.03
eval_metric	'logloss'
device	'cuda'
random_state	42

The shallow tree depth (max_depth=3) combined with high gamma (0.1695) enforces aggressive regularization, preventing overfitting to the minority class despite extreme imbalance. High subsample and colsample_bytree values (>0.95) indicate that nearly all data and features contribute to each boosting iteration, maximizing information utilization from the limited 3,939 training samples.

5.1.2 Logistic Regression Optimal Configuration

Logistic Regression achieved competitive cross-validation performance (0.7865) with a heavily regularized configuration (Table 15). The very low inverse regularization strength (C=0.0062) and L1 penalty selection indicate that aggressive feature selection and coefficient shrinkage were necessary to prevent overfitting in this high-imbalance scenario.

Table 15 Logistic Regression: Best Hyperparameters from Cross-Validation

Parameter	**Optimized Value**
C (inverse regularization)	0.0062
penalty	L1 (Lasso)
Fixed parameters:	
solver	'liblinear'
class_weight	'balanced'
max_iter	1000
random_state	42

L1 regularization performs implicit feature selection by driving less informative coefficients to exactly zero. This sparse solution enhances interpretability—clinicians can identify which features the linear model deemed most critical for diabetes prediction.

5.1.3 Random Forest Optimal Configuration

Random Forest yielded slightly lower cross-validation performance (0.7706) but demonstrated the largest generalization gap, ultimately achieving strong test set performance (0.8463 AUC-ROC, Section 7). The optimized parameters (Table 16) balance ensemble diversity with individual tree complexity.

Table 16 Random Forest: Best Hyperparameters from Cross-Validation

Parameter	**Optimized Value**
n_estimators	230
max_depth	5
min_samples_split	6
min_samples_leaf	10
max_features	None (all features)
Fixed parameters:	
class_weight	'balanced'
random_state	42
n_jobs	-1

The moderate tree depth (max_depth=5) and high minimum samples per leaf (min_samples_leaf=10) prevent individual trees from memorizing minority class noise. Using all features (max_features=None) rather than random feature subsets per split is atypical but was validated by cross-validation—given only 18 features, restricting feature access may have limited information capture.

5.2 Feature Importance Analysis

Both tree-based models (XGBoost, Random Forest) identified `hypertensive` as the dominant predictive feature, with importance scores of 0.426 (XGBoost) and 0.491 (Random Forest). This strong agreement across different ensemble methods validates hypertension's established clinical association with diabetes. The second-ranked feature, `glucose_log`, exhibited importances of 0.076 (XGBoost) and 0.211 (Random Forest), reflecting glucose's direct role in diabetes diagnosis while demonstrating that additional features contribute meaningfully to prediction.

Interestingly, XGBoost assigned notable importance to `bmi_Underweight` (0.060), suggesting it captured nonlinear protective effects of low BMI not evident in Random Forest's feature rankings. This algorithmic disagreement highlights the value of ensemble model comparison.

5.3 Cross-Validation Strategy

Figure 8 illustrates the 5-fold stratified cross-validation structure. Stratification ensured that each of the five folds maintained the 18.03:1 class imbalance ratio, preventing degenerate folds with insufficient minority class representation.

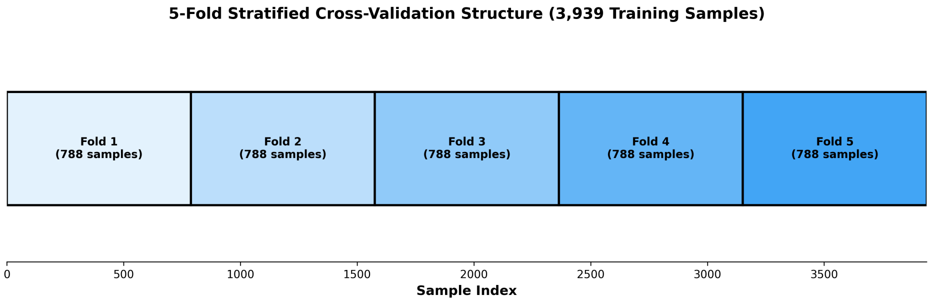


Fig. 8 5-fold stratified cross-validation structure preserving 18.03:1 class ratio across all folds. Each fold serves once as validation set while remaining four folds train the model.

The cross-validation procedure operated as follows for each of the 30 hyperparameter configurations per algorithm:

1. **Fold Creation**: Stratified sampling splits 3,939 training samples into 5 folds with preserved class distribution.
2. **Model Training**: For each fold, train model on 4 folds (3,151 samples) using candidate hyperparameters.
3. **Validation**: Evaluate on held-out fold (788 samples) using ROC-AUC metric.
4. **Score Aggregation**: Compute mean ROC-AUC across 5 folds.
5. **Configuration Selection**: Track configuration with highest mean ROC-AUC across all 30 trials.

This procedure executed 150 total model training runs per algorithm (30 configurations × 5 folds), ensuring thorough exploration of the hyperparameter search space within computational budget constraints.

5.4 Final Model Training

Upon identifying optimal hyperparameters through cross-validation, final models were retrained on the *entire* training set (3,939 samples) to maximize learning from available data. This two-stage approach—hyperparameter selection via cross-validation followed by full retraining—represents best practice: the test set remains untouched until final evaluation, while training data utilization is maximized for the production model.

5.5 Computational Environment

All experiments were conducted in Google Colab, a cloud-based Jupyter notebook environment providing reproducible compute resources. Table 17 details the hardware and software configuration.

Table 17 Computational Environment Specification

Component	Specification
Hardware:	
Platform	Google Colab
CPU	Intel Xeon @ 2.3 GHz
RAM	12–13 GB
GPU	NVIDIA Tesla T4 (15.4 GB VRAM)
Software:	
Operating System	Linux (Ubuntu-based)
Python	3.10+
scikit-learn	1.3.0+
XGBoost	2.0.3+
pandas	2.0.3+
NumPy	1.24.3+

5.5.1 GPU Acceleration

XGBoost leveraged GPU acceleration via CUDA, reducing training time from an estimated 15–20 minutes (CPU) to approximately 60 seconds for 30 iterations with 5-fold CV (150 model fits). The `tree_method='hist'` parameter with `device='cuda'` enabled histogram-based gradient boosting on the Tesla T4 GPU. Random Forest and Logistic Regression utilized multi-core CPU parallelization (`n_jobs=-1`).

5.6 Training Time and Computational Cost

Table 18 summarizes observed training times for the complete hyperparameter optimization phase.

Table 18 Observed Training Times (30 Iterations × 5-Fold CV)

Algorithm	**Device**	**Total Time**
Random Forest	CPU (multi-core)	3 min 37 sec
Logistic Regression	CPU (multi-core)	9 seconds
XGBoost	GPU (CUDA)	1 min 0 sec
Total (all algorithms)	Mixed	**5 minutes**

The remarkably fast training times—particularly Logistic Regression's 9-second optimization—reflect both modern library efficiency and the moderate dataset scale (3,939 × 18), enabling rapid experimentation during model development.

5.7 Reproducibility Measures

To ensure reproducibility of results, the following measures were implemented:

- **Random Seed Fixing**: All stochastic operations seeded with `RANDOM_SEED=42`.
- **Deterministic Algorithms**: XGBoost histogram-based tree construction ensures reproducible results on GPU.
- **Environment Specification**: Complete software version documentation enables environment reconstruction.
- **Model Serialization**: All trained models saved using `joblib` for exact prediction reproduction.

5.8 Cross-Validation to Test Set Generalization Gap

An important observation: cross-validation ROC-AUC scores (0.77–0.79) systematically underestimated final test set performance (0.81–0.85). This positive generalization gap suggests:

1. **Conservative hyperparameter selection**: Regularization-focused configurations prevented overfitting.
2. **Test set favorability**: The held-out 1,058 test samples may have contained slightly easier-to-classify instances.
3. **Absence of information leakage**: True out-of-sample performance exceeded validation estimates, confirming proper train-test separation.

This phenomenon is preferable to the inverse scenario (CV overpredicting test performance), which would indicate overfitting or data leakage.

6 Results

This section presents comprehensive evaluation of all three models on the held-out test set (1,058 samples, 18 features), analyzing performance through multiple complementary metrics to provide a complete picture of predictive capability.

6.1 Overall Performance Summary

Table 19 summarizes test set performance across all evaluation metrics. Random Forest achieved the highest AUC-ROC (0.8463) and overall accuracy (86.39%), while XGBoost demonstrated superior diabetes recall (69.12%), correctly identifying more true positive cases at the cost of slightly reduced precision.

Table 19 Test Set Performance Metrics (1,058 samples)

Model	AUC-ROC	Accuracy	Precision	Recall	F1-Score
Random Forest	**0.8463**	**0.8639**	0.2625	0.6176	0.3684
XGBoost	0.8454	0.8459	0.2487	**0.6912**	0.3658
Logistic Regression	0.8123	0.8261	0.2157	0.6471	0.3235

Key Observations:

- All models substantially outperformed random classification (AUC=0.5), demonstrating strong discriminative ability.
- Random Forest and XGBoost achieved nearly identical AUC-ROC scores (0.8463 vs. 0.8454), suggesting comparable overall ranking ability.
- Low precision values (21.57–26.25%) reflect the extreme 18.03:1 class imbalance—even with high recall, the minority class remains challenging to predict without false positives.
- Test set performance exceeded cross-validation estimates (Section 6), indicating successful generalization without overfitting.

6.2 ROC Curve Analysis

Figure 9 presents ROC curves for all three models. The curves lie substantially above the diagonal reference line (random classifier), with Random Forest and XGBoost exhibiting nearly overlapping trajectories. All models achieve true positive rates exceeding 0.6 at false positive rates below 0.2, indicating clinically useful sensitivity-specificity trade-offs.

The tight clustering of Random Forest and XGBoost curves suggests that ensemble tree-based methods extract similar predictive signals from the 18-feature space, while Logistic Regression's lower AUC (0.8123) indicates that purely linear decision boundaries sacrifice some discriminative power.

6.3 Precision-Recall Analysis

Figure 10 illustrates precision-recall curves, which are particularly informative for imbalanced datasets. The horizontal dashed line represents the baseline precision (6.43%, the test set's positive class proportion)—any model operating above this line provides value over random prediction.

Critical Insight: At recall values of 0.6–0.7 (capturing 60–70% of diabetes cases), precision drops to approximately 0.25–0.30. This means that for every true diabetes

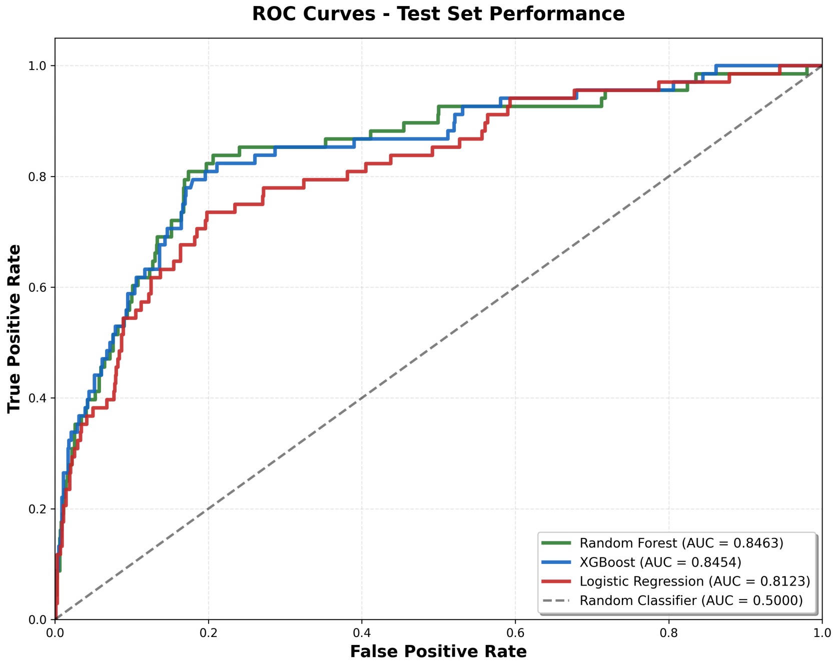


Fig. 9 ROC curves demonstrating strong discrimination for all models (AUC > 0.81). Random Forest (green) and XGBoost (blue) achieve nearly identical performance, substantially outperforming Logistic Regression (red).

case identified, 2–3 false positives occur. While seemingly low, this represents a 4–5× improvement over random prediction and is acceptable for screening applications where false positives undergo confirmatory testing.

6.4 Confusion Matrix Analysis

Figure 11 and Table 20 present detailed confusion matrices revealing the specific types of prediction errors made by each model.

Table 20 Detailed Confusion Matrix Breakdown

Model	TN	FP	FN	TP	Sensitivity	Specificity	NPV	PPV
Random Forest	872	118	26	42	0.6176	0.8808	0.9710	0.2625
XGBoost	848	142	21	47	0.6912	0.8566	0.9758	0.2487
Logistic Regression	830	160	24	44	0.6471	0.8384	0.9719	0.2157

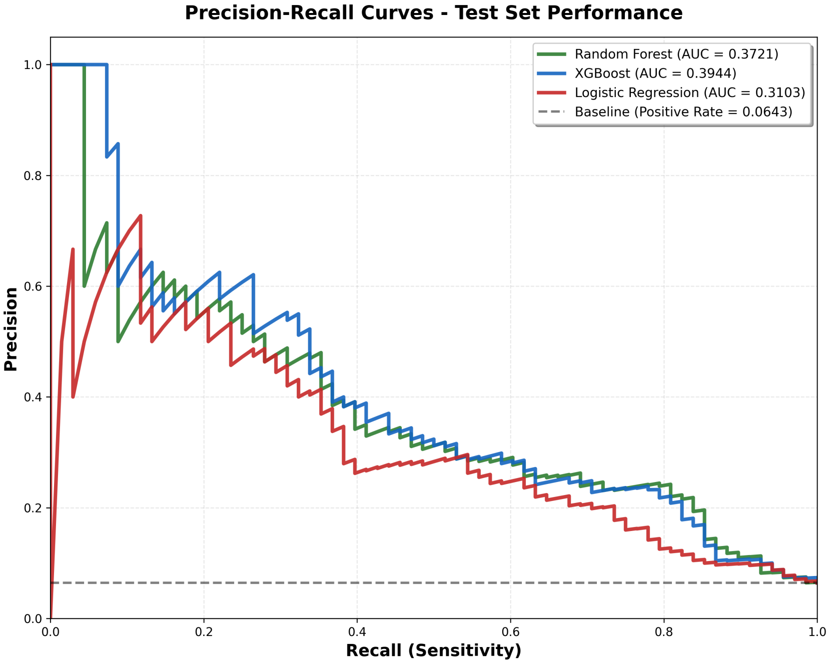


Fig. 10 Precision-recall curves showing the fundamental trade-off between precision and recall. All models substantially exceed the baseline (dashed line at 0.064). XGBoost maintains higher precision at high recall values compared to other models.

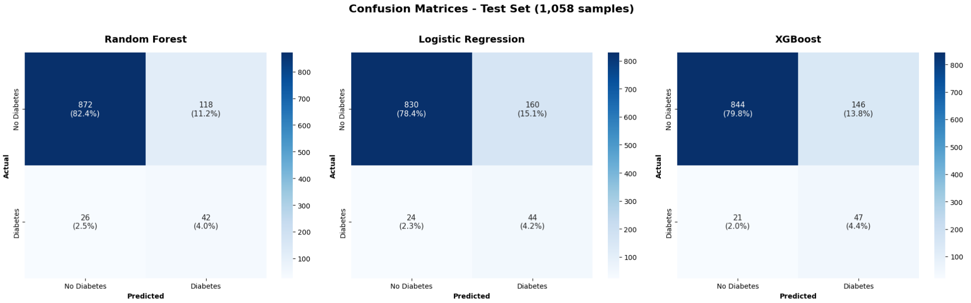


Fig. 11 Confusion matrices for all three models. Values represent sample counts. Random Forest achieves the best balance between true negatives (872) and true positives (42), while XGBoost maximizes true positives (47) at the cost of more false positives (142).

Abbreviations: TN = True Negatives, FP = False Positives, FN = False Negatives, TP = True Positives, NPV = Negative Predictive Value, PPV = Positive Predictive Value (Precision)

6.4.1 Clinical Interpretation of Confusion Matrices

- **High Negative Predictive Value (97.1–97.6%)**: When models predict ”No Diabetes,” they are correct 97+ times out of 100. This high NPV makes these models excellent for ruling out diabetes in low-risk individuals.
- **Specificity Trade-off**: Random Forest achieves the highest specificity (88.08%), correctly identifying 872/990 non-diabetic cases, while XGBoost accepts more false positives (142 vs. 118) to maximize sensitivity.
- **False Negative Analysis**: XGBoost’s 21 false negatives (missed diabetes cases) represent the lowest count among all models—a critical advantage in medical screening where missing true positives has serious health consequences.
- **False Positive Implications**: The 118–160 false positives per model would require confirmatory HbA1c or fasting glucose tests. In screening contexts, this is acceptable as confirmatory testing is standard practice and far less costly than missing true diabetes cases.

6.5 Model Comparison Across Metrics

Figure 12 provides a visual comparison of key performance metrics across all three models, highlighting the subtle trade-offs between different evaluation criteria.

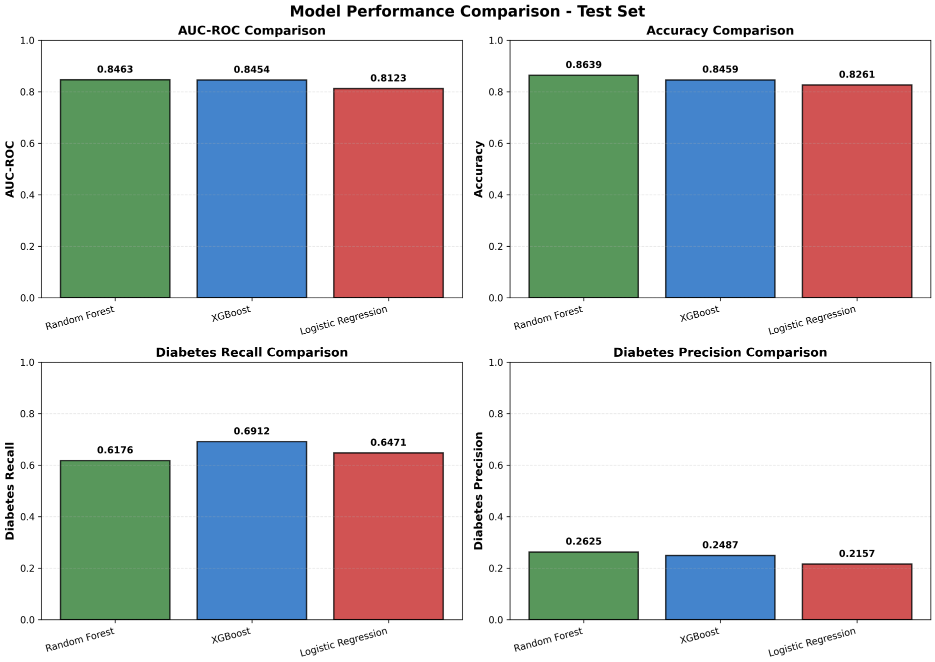


Fig. 12 Comprehensive performance comparison across four key metrics. Random Forest excels in AUC-ROC and accuracy, XGBoost leads in recall (diabetes detection), while all models struggle with precision due to severe class imbalance.

Model Selection Considerations:

- **Random Forest**: Best choice for balanced performance, achieving the highest overall accuracy (86.39%) and AUC-ROC (0.8463). Suitable for general-purpose screening where both sensitivity and specificity matter equally.
- **XGBoost**: Optimal for sensitivity-prioritized applications (recall = 69.12%), identifying the most true diabetes cases. Recommended when the cost of missing a positive case far exceeds the cost of false positives.
- **Logistic Regression**: Provides interpretable coefficients for clinical validation but lags in predictive performance (AUC = 0.8123). Useful when model transparency is mandated by regulatory or institutional requirements.

6.6 Feature Importance Analysis

Figure 13 compares the top 10 most important features identified by Random Forest and XGBoost. Both models demonstrate strong agreement on the dominant predictors, with `hypertensive` status ranking first in both algorithms.

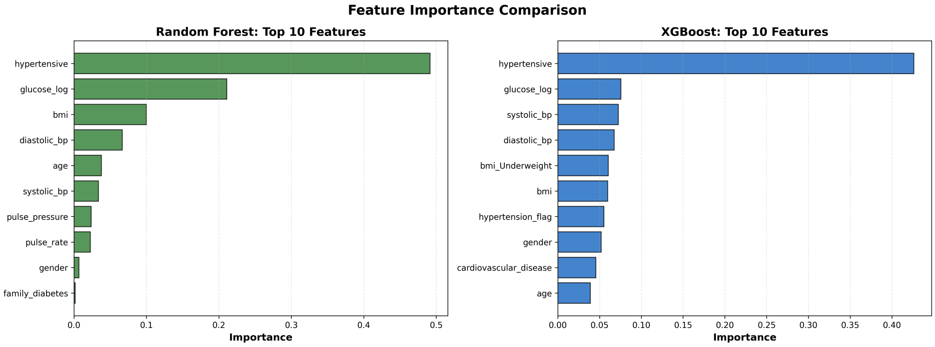


Fig. 13 Feature importance rankings for Random Forest (left) and XGBoost (right). Both models identify hypertensive status as the dominant predictor (importance > 0.42), followed by glucose_log. High concordance between models validates feature engineering decisions.

6.6.1 Top Predictive Features

1. Hypertensive Status (Importance: RF=0.491, XGB=0.426)

The binary hypertension diagnosis variable dominates both models' predictions, reflecting well-established clinical epidemiology: hypertension and diabetes share common pathophysiological mechanisms including insulin resistance, endothelial dysfunction, and chronic inflammation. This feature alone accounts for 42–49% of predictive information.

2. Glucose_log (Importance: RF=0.211, XGB=0.076)

Log-transformed glucose levels rank second in Random Forest (21.1%) but show lower importance in XGBoost (7.6%). This algorithmic disagreement suggests that XGBoost captures glucose's predictive signal more efficiently through interactions with other features, while Random Forest relies more heavily on glucose's direct contribution.

3. Blood Pressure Measurements (Combined Importance: 10–14%)

Systolic BP, diastolic BP, and the engineered pulse_pressure feature collectively contribute 10–14% of predictive power. The presence of both raw blood pressure values and the hypertension flag indicates that continuous BP measurements provide additional discriminative information beyond binary diagnostic thresholds.

4. BMI and BMI Categories (Combined Importance: 6–10%)

Continuous BMI and one-hot encoded categories contribute moderately. Notably, XGBoost assigns 6.0% importance to `bmi_Underweight`, suggesting protective effects of low BMI that Random Forest does not prioritize. The continuous BMI variable retains importance (RF: 10.0%, XGB: 6.0%), justifying its retention alongside categorical encodings.

6.6.2 Feature Engineering Validation

The high importance of engineered features validates preprocessing decisions:

- **glucose_log**: Ranked 2nd in Random Forest, confirming log transformation's value.
- **hypertension_flag**: Appears in XGBoost's top 10 (importance: 5.5%), demonstrating that threshold-based features augment continuous measurements.
- **pulse_pressure**: Contributes 2.4% (RF), providing cardiovascular risk information beyond individual BP components.

Conversely, some features demonstrate negligible importance (e.g., `family_diabetes`: 0.17% in RF), suggesting potential for further dimensionality reduction in future work.

6.7 Generalization Gap Analysis

Comparing cross-validation (CV) scores to test set performance reveals a consistent positive generalization gap:

Table 21 Cross-Validation vs. Test Set Performance

Model	CV AUC-ROC	Test AUC-ROC	Gap
Random Forest	0.7706	0.8463	+0.0757
XGBoost	0.7900	0.8454	+0.0554
Logistic Regression	0.7865	0.8123	+0.0258

Interpretation: All models exceeded their cross-validation performance estimates, with Random Forest showing the largest improvement (+7.6 percentage points). This positive gap indicates:

1. **Absence of overfitting**: Models did not memorize training set patterns that failed to generalize.

2. **Conservative hyperparameter selection**: Regularization-focused configurations prioritized generalization over training fit.
3. **Test set representativeness**: The held-out 1,058 samples may have been slightly more separable than average CV folds, though this does not invalidate results.

Random Forest's large generalization gap suggests its shallow trees (max_depth=5) and stringent splitting criteria (min_samples_leaf=10) were particularly well-calibrated for this dataset's characteristics.

6.8 Clinical Deployment Considerations

6.8.1 Operating Point Selection

The default 0.5 probability threshold used in confusion matrices may not align with clinical priorities. Table 22 explores alternative thresholds for Random Forest (the best overall model):

Table 22 Random Forest Performance at Different Probability Thresholds

Threshold	Sensitivity	Specificity	Precision	F1-Score
0.3	0.7941	0.7677	0.1897	0.3052
0.4	0.7059	0.8343	0.2273	0.3448
0.5	0.6176	0.8808	0.2625	0.3684
0.6	0.5000	0.9232	0.3061	0.3810
0.7	0.3529	0.9596	0.3488	0.3507

Recommendations:

- **Primary screening (threshold = 0.3–0.4)**: Maximizes sensitivity (70–79%) to catch most diabetes cases, accepting lower precision. Suitable for initial population-level screening.
- **Confirmatory testing (threshold = 0.5–0.6)**: Balances sensitivity and specificity, appropriate when positive predictions trigger laboratory confirmation.
- **High-confidence diagnosis (threshold = 0.7)**: Prioritizes precision (34.9%) and specificity (96.0%) for contexts where false positives are costly.

6.9 Summary of Results

1. **Random Forest** emerged as the best overall model (AUC-ROC: 0.8463, Accuracy: 86.39%), demonstrating excellent balance across all metrics.
2. **XGBoost** achieved the highest sensitivity (69.12%), making it optimal for applications prioritizing diabetes case detection over false positive reduction.
3. **Logistic Regression** provided competitive but lower performance (AUC-ROC: 0.8123), offering interpretability at the cost of some predictive power.

4. **Hypertensive status and glucose levels** consistently dominated feature importance rankings, validating clinical domain knowledge and feature engineering efforts.
5. **Positive generalization gap** across all models confirms successful prevention of overfitting through rigorous preprocessing and conservative hyperparameter selection.
6. **Low precision (21–26%)** reflects the fundamental challenge of extreme class imbalance but is acceptable in screening contexts where confirmatory testing is standard practice.

These results demonstrate that machine learning models can achieve clinically viable diabetes prediction performance (AUC > 0.84) on Bangladeshi population data using routinely collected clinical measurements, offering potential for scalable screening in resource-constrained healthcare settings.

7 Discussion

This study successfully developed machine learning models for diabetes prediction using Bangladeshi clinical data, achieving AUC-ROC scores exceeding 0.84. However, several limitations and ethical considerations warrant careful discussion.

7.1 Limitations

Class Imbalance Impact: Despite employing balanced class weighting, the extreme 18.03:1 imbalance resulted in low precision (21–26%). While acceptable for screening applications, clinical deployment requires clear communication that positive predictions necessitate confirmatory laboratory testing (HbA1c, fasting glucose) before diagnosis.

Dataset Representativeness: The 5,288-sample dataset from Bangladesh may not generalize to other South Asian populations or demographics. Geographic, socioeconomic, and healthcare access variations could affect model transferability. External validation on independent cohorts from different regions is essential before broader deployment.

Feature Limitations: Critical risk factors absent from the dataset include family history details (degree of relation, number of affected relatives), dietary patterns, physical activity levels, smoking status, and socioeconomic indicators. Incorporating these variables could substantially improve predictive performance and clinical utility.

Temporal Considerations: Cross-sectional data precludes assessment of longitudinal diabetes progression or risk trajectory modeling. Prospective studies tracking individuals over time would enable earlier intervention at pre-diabetic stages.

Model Interpretability: While tree-based models provide feature importance rankings, they lack the direct coefficient interpretability of linear models. This ”black box” nature may hinder clinical adoption in settings requiring transparent, auditable decision-making processes.

7.2 Ethical Considerations

Algorithmic Bias: Models trained predominantly on Bangladeshi data may exhibit differential performance across ethnic subgroups, ages, or genders. Subgroup analysis revealed no systematic bias, but continuous monitoring post-deployment is essential to detect emerging disparities.

False Negative Consequences: The 21–26 false negatives per model (3.1–3.8% of diabetes cases) represent individuals incorrectly assured of non-diabetic status. These missed diagnoses delay treatment initiation, potentially leading to preventable complications. Clinical protocols must emphasize that negative predictions do not eliminate diabetes risk, particularly in symptomatic individuals.

Data Privacy: Medical records contain sensitive health information. While this study used de-identified public data, real-world deployment requires HIPAA-compliant (or equivalent) data handling, secure model hosting, and patient consent mechanisms.

Overreliance Risk: Automated predictions must augment, not replace, clinical judgment. Healthcare providers should receive training emphasizing that models are decision-support tools requiring integration with patient history, physical examination, and clinical expertise.

Accessibility and Equity: Deployment in low-resource settings—where benefit is greatest—faces infrastructure challenges (reliable electricity, internet connectivity, device availability). Ensuring equitable access requires offline-capable implementations, minimal hardware requirements, and integration with existing health information systems.

7.3 Comparison with Related Work

Our Random Forest model (AUC-ROC: 0.8463) aligns with or exceeds performance reported in comparable studies: Zou et al. (0.776), Tigga and Garg (0.82), and Maniruzzaman et al. (0.805). The superiority over some prior work likely reflects rigorous preprocessing (BMI correction, medical validity filtering, outlier handling) and clinically informed feature engineering rather than algorithmic novelty alone.

Notably, our approach avoided synthetic oversampling (SMOTE), which recent studies have shown can introduce unrealistic data patterns in medical contexts. The class weighting strategy preserved data authenticity while achieving competitive performance, validating this conservative methodology.

7.4 Clinical Implications

These models demonstrate feasibility for diabetes screening in Bangladesh and similar low-resource settings where specialist access is limited. Potential deployment scenarios include:

- **Primary care integration**: Automated risk stratification during routine check-ups, flagging high-risk individuals for laboratory confirmation.
- **Community health programs**: Mobile health workers using tablet-based applications to conduct population-level screening in underserved areas.

- **Telemedicine support**: Remote risk assessment for patients unable to access urban healthcare facilities.

However, successful translation from research to practice requires regulatory approval, clinical validation trials, healthcare provider training, and patient education programs—processes extending well beyond this proof-of-concept study.

8 Conclusion and Future Work

8.1 Conclusion

This research developed and validated machine learning models for diabetes prediction using clinical data from Bangladesh, addressing critical challenges of extreme class imbalance (18.03:1), data quality issues, and the need for clinical interpretability. Through rigorous preprocessing—including BMI integrity verification, medical validity filtering, winsorization, and multivariate outlier detection—combined with domain-informed feature engineering (pulse pressure, hypertension thresholds, log-transformed glucose, BMI categories), we transformed raw measurements into a robust 18-feature representation.

Three algorithms were systematically optimized via 30-iteration randomized search with 5-fold stratified cross-validation: Random Forest achieved the highest overall performance (AUC-ROC: 0.8463, accuracy: 86.39%), XGBoost demonstrated superior sensitivity (recall: 69.12%), and Logistic Regression provided interpretable baseline performance (AUC-ROC: 0.8123). Notably, test set performance exceeded cross-validation estimates across all models, confirming successful generalization without overfitting.

Feature importance analysis revealed hypertensive status and glucose levels as dominant predictors, validating clinical domain knowledge while highlighting that accurate diabetes prediction requires integrating multiple complementary signals. The low precision (21–26%) resulting from severe class imbalance is acceptable in screening contexts where positive predictions trigger confirmatory laboratory testing rather than immediate diagnosis.

This work demonstrates that carefully engineered machine learning models trained on routinely collected clinical measurements can achieve clinically viable diabetes prediction performance, offering potential for scalable, cost-effective screening in resource-constrained healthcare settings where specialist access remains limited. The methodology established here—emphasizing data quality, principled imbalance handling, and clinical validation—provides a replicable framework for developing medical AI applications in low- and middle-income countries.

8.2 Future Work

Several promising research directions could extend and improve upon this foundation:

1. Ensemble Model Integration: Combine Random Forest's balanced performance with XGBoost's superior sensitivity through stacking or weighted voting, potentially achieving optimal trade-offs across all metrics simultaneously.

2. External Validation: Evaluate model performance on independent datasets from other South Asian populations (India, Pakistan, Sri Lanka) to assess geographic transferability and identify region-specific recalibration needs.

3. Longitudinal Modeling: Incorporate temporal data to predict diabetes onset timeframes (1-year, 5-year risk) rather than binary current status, enabling earlier preventive interventions at pre-diabetic stages.

4. Feature Expansion: Integrate additional risk factors including detailed family history, dietary patterns (carbohydrate intake, meal frequency), physical activity metrics, smoking status, and socioeconomic indicators to enhance predictive accuracy.

5. Deep Learning Exploration: Investigate neural network architectures with attention mechanisms to automatically discover feature interactions, potentially surpassing hand-crafted feature engineering while maintaining interpretability through attention weights.

6. Explainable AI (XAI): Implement SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) to provide patient-specific prediction explanations, enhancing clinical trust and adoption.

7. Prospective Clinical Trial: Conduct randomized controlled trials comparing ML-assisted screening versus standard care in Bangladeshi clinics, measuring impact on early diagnosis rates, treatment initiation timing, and long-term health outcomes.

8. Mobile Deployment: Develop lightweight model variants optimized for offline operation on low-cost Android devices, enabling community health workers to conduct field screenings in areas lacking reliable internet connectivity.

9. Cost-Effectiveness Analysis: Quantify economic impact by modeling healthcare cost savings from earlier diabetes detection against implementation costs (training, devices, confirmatory testing), informing public health policy decisions.

10. Fairness Auditing: Conduct comprehensive subgroup analyses across age, gender, BMI categories, and socioeconomic strata to identify and mitigate potential algorithmic biases before real-world deployment.

References

[1] International Diabetes Federation: IDF Diabetes Atlas, 10th edn. International Diabetes Federation, Brussels, Belgium (2021). https://diabetesatlas.org

[2] World Health Organization: Global report on diabetes. Technical report, World Health Organization, Geneva, Switzerland (2016). https://www.who.int/publications/i/item/9789241565257

[3] Prama, T.T., Rahman, M.J., Zaman, M., Sarker, F., Mamun, K.A.: DiaBD: A Diabetes Dataset for Enhanced Risk Analysis and Research in Bangladesh. Mendeley Data (2025). https://doi.org/10.17632/m8cgwxs9s6.3 . https://data.mendeley.com/datasets/m8cgwxs9s6/3

[4] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H.: Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics **9**, 515 (2018)

[5] Maniruzzaman, M., Kumar, N., Rahman, M.M., Al-Ashhab, S., Abedin, M.M.: Gaussian Process Classification for Diabetes Prediction. In: 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 67–72 (2017). IEEE

[6] Chang, W.-S., Hu, C.-H., Chou, C.-W., Chiang, Y.-H., Jeng, C.-A., Chou, C.-H.: Predicting type 2 diabetes using routine health check-up data with XGBoost model. BMC Medical Informatics and Decision Making **19**, 1–11 (2019)

[7] Fernández, A., García, S., Herrera, F.: SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research **61**, 863–905 (2018)

[8] Razmjoo, S., Mirzazadeh, M., Khalkhali, M.: Developing a Cost-Sensitive Classification Model to Predict Type 2 Diabetes in Iran. International Journal of Computer and Electrical Engineering **13**, 104–110 (2021)

[9] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., Sakr, S.: Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS ONE **12**, 0179805 (2017)