Enhancing Healthcare: Machine Learning for Diabetes Prediction and Retinopathy Risk Evaluation

Ghinwa Barakat[1], Samer El Hajj Hassan[2]*, Nghia Duong-Trung[3], Wiam Ramadan[4]
Biological and Chemical Sciences Department, Lebanese International University, Beirut, Lebanon[1]
Computer Science Department, International University of Applied Sciences, Berlin, Germany[2]
German Research Centre for Artificial Intelligence (DFKI), Berlin, Germany[3]
Nutrition and Food Science Department, Lebanese International University, Beirut, Lebanon[4]

***Abstract*—Diabetes mellitus stands as a major public health issue that affects millions globally. Among the various complications associated with diabetes, diabetic retinopathy presents a significant concern, affecting approximately one-third of diabetic patients. Early detection of diabetic retinopathy is paramount, as timely treatment can significantly reduce the risk of severe visual impairment. The study employs advanced machine learning techniques to predict diabetes and assess risk levels for retinopathy, aiming to enhance predictive accuracy and risk stratification in clinical settings. This approach contributes to better management and treatment outcomes. A diverse array of machine learning models including Logistic Regression, Random Forest, XGBoost, voting classifiers was used. These models were applied to a meticulously selected dataset, specifically designed to include comprehensive diabetic indicators along with retinopathy outcomes, enabling a detailed comparative analysis. Among the evaluated models, XGBoost demonstrated superior performance in terms of accuracy, sensitivity, and computational efficiency. This model excelled in identifying risk levels among diabetic patients, providing a reliable tool for early detection of potential retinopathy. The findings suggest that the integration of machine learning models, particularly XGBoost, into the healthcare system could significantly enhance early screening and personalized treatment plans for diabetic retinopathy. This advancement holds the potential to improve patient outcomes through timely and accurate risk assessment, paving the way for targeted interventions.**

***Keywords*—Machine learning; diabetes prediction; artificial intelligence in healthcare; XGBoost; Random Forest**

I. Introduction

A. Diabetes Mellitus

Diabetes mellitus (DM) is a complex metabolic disorder categorized by raised blood glucose levels, resulting from defects in insulin secretion, insulin action, or both. This condition represents a key health concern worldwide, affecting millions of individuals and imposing an extensive economic problem on healthcare systems [1]. The occurrence of diabetes has been progressively rising, fueled by sedentary lifestyles, poor dietary habits, obesity, ethnicity, advancing age and genetic predisposition. Type 1 diabetes mellitus (T1DM) is characterized by autoimmune destruction of pancreatic beta cells, leading to absolute insulin deficiency [2]. T1DM often develops early in life, although it can occur at any age. Individuals with T1DM require lifelong insulin therapy to survive. Type 2 diabetes mellitus (T2DM), the most prevalent form accounting for most cases worldwide, typically arises from a combination of insulin resistance and inadequate insulin secretion [3]. It typically develops in adulthood, although there has been a concerning rise in its occurrence among children and adolescents due to the increasing prevalence of obesity and sedentary lifestyles [4]. In T2DM, the body becomes resistant to the action of insulin, and the pancreas may fail to produce enough insulin to compensate for this resistance. This results in elevated blood glucose levels. While genetic factors play a role in predisposing individuals to T2DM, lifestyle factors such as poor diet, lack of physical activity, and obesity are significant contributors to its development [5]. Gestational diabetes (GDM) occurs during pregnancy and is associated with increased risk of both maternal and fetal complications. GDM poses risks to both the mother and the fetus, including an increased likelihood of complications such as macrosomia (large birth weight), birth trauma, hypoglycemia in the newborn, and an elevated risk of developing type 2 diabetes for both the mother and child later in life [6]. While GDM typically resolves after childbirth, affected women are at an increased risk of developing T2DM in the future.

The economic impact of diabetes spans healthcare costs, productivity losses, and societal implications. Direct healthcare expenditures include medication, hospitalizations, and complications management. Additionally, indirect costs arise from productivity declines due to disability, absenteeism, and premature mortality [7]. The socioeconomic consequences extend to reduced quality of life, disparities in healthcare access, and strained healthcare systems. Addressing this global health challenge requires a multifaceted approach encompassing prevention strategies, early detection, lifestyle modifications, access to healthcare services, and effective management and treatment options.

B. Diabetes and Retinopathy

In addition to its metabolic manifestations, diabetes predisposes individuals to numerous complications, including cardiovascular disease, neuropathy, nephropathy, and retinopathy. Among these, diabetic retinopathy (DR) stands out as a significant cause of preventable blindness, highlighting the importance of understanding its pathogenesis and management. This condition affects the eyes, specifically the retina, the light-sensitive tissue at the back of the eye. It is a microvascular complication of diabetes that affects the retinal vasculature, leading to progressive damage and vision loss [8]. DR, whose incidence is high in the working-age population, prevails all over

the world and is estimated to reach 191 million cases by 2030 [9, 10]. The pathogenesis of DR is multifactorial, involving chronic hyperglycemia, oxidative stress, inflammation, and vascular dysfunction [11]. Hyperglycemia-induced metabolic abnormalities contribute to the development of microaneurysms, capillary nonperfusion, and increased vascular permeability, culminating in retinal ischemia and neovascularization. Chronic inflammation further exacerbates vascular damage and promotes the release of angiogenic factors, perpetuating a vicious cycle of retinal injury. It progresses through several stages, starting with non-proliferative diabetic retinopathy (NPDR), where small blood vessels in the retina weaken and leak fluid into the surrounding tissue, causing swelling and leading to blurry vision. As the disease advances, it can enter the proliferative stage, characterized by the growth of abnormal blood vessels on the surface of the retina. These vessels are fragile and prone to bleeding, leading to further vision impairment and, in severe cases, retinal detachment [11]. Additionally, diabetic macular edema (DME) can occur, where fluid accumulates in the macula, the central part of the retina responsible for sharp, central vision, leading to significant vision loss. These changes can impair vision and, if left untreated, result in blindness. Symptoms may not be noticeable in the early stages, but as the condition progresses, individuals may experience blurred vision, floaters, and even complete vision loss. The risk factors for diabetic retinopathy include the duration of diabetes, poorly controlled blood sugar levels, high blood pressure, and high cholesterol [11]. Early detection and timely intervention are crucial for preventing vision loss in DR. Several diagnostic modalities are available for the assessment of DR, including dilated fundus examination, fundus photography, optical coherence tomography (OCT), and fluorescein angiography.

The management of diabetic retinopathy is multifaceted and involves lifestyle modifications, optimizing glycemic control, blood pressure management, and lipid-lowering therapy to reduce systemic risk factors. Patient education and regular ophthalmic screenings are essential components of comprehensive diabetes care to minimize the impact of retinopathy on visual function. Collaborative care between endocrinologists, ophthalmologists, and other healthcare providers is essential to provide comprehensive management and minimize the impact of this potentially sight-threatening complication of diabetes.

C. Machine Learning Significance

Machine learning (ML) incorporates a suite of computational techniques that enable systems to learn from and make predictions or decisions based on data. In predictive modeling, ML algorithms use historical data as input to predict new output values [12]. These models iteratively learn from the data, improving their accuracy over time without being explicitly programmed to perform specific tasks. This capability makes ML an invaluable tool across various domains, including finance, marketing, and notably, healthcare.

D. Predictive Modeling Importance

In the realm of healthcare, predictive analytics plays a pivotal role, especially in the early detection of diseases and the stratification of patient risk levels. For chronic conditions like diabetes mellitus [13], early prediction and diagnosis can significantly improve patient outcomes and reduce healthcare costs. According to the International Diabetes Federation, approximately 537 million adults (20-79 years) were living with diabetes in 2021, and this number is expected to rise to 643 million by 2030 and 783 million by 2045 [14]. Specifically, in the context of diabetic retinopathy [15], Diabetic retinopathy is a leading cause of blindness in working-age adults, and early detection and management are crucial to prevent vision loss [16]. However, current methods for predicting diabetes and evaluating the risk of retinopathy often rely on traditional statistical models, which may not capture the complex relationships among various risk factors. Machine learning has emerged as a powerful tool in healthcare, offering advanced methods for predicting and diagnosing diseases by analyzing large datasets and identifying patterns that may not be apparent with traditional methods [17]. Machine learning models can analyze complex datasets to predict disease onset and progression, enabling healthcare providers to prioritize patients with a high risk of vision loss for early treatment, thereby optimizing resource allocation and improving patient outcomes [17]. Recent studies highlight the importance and advancements in healthcare predictive models, such as Darmadi *et al.* (2023) [18] who enhanced global health system resilience post-COVID-19 through grounded theory approaches, Lampezhev *et al.* (2022) [19] who developed methods for analyzing the uniqueness of personal medical data, and Muthaiyah *et al.* (2023) [20] who presented a binary survivability prediction classification model for osteosarcoma prognosis. These studies underline the critical role of advanced machine learning techniques in modern healthcare. Additionally, Duong-Trung et al. (2019) [21] proposed a workflow for medical diagnosis through the lens of the machine learning perspective, emphasizing the integration of machine learning to boost automatic medical decision-making and reduce data overload.

By leveraging machine learning algorithms, this study aims to enhance the accuracy and reliability of diabetes mellitus prediction and retinopathy risk evaluation, ultimately improving patient outcomes and reducing healthcare costs.

This study introduces a novel machine learning-based approach for predicting diabetes and evaluating the risk of diabetic retinopathy. This research integrates multiple advanced machine learning algorithms, including XGBoost, to enhance predictive accuracy.

II. Aim of the Study

This study aims to harness the power of machine learning to enhance the prediction and risk assessment capabilities for diabetes and its consequential complication, diabetic retinopathy. The primary objectives of this research are:

1) To develop and implement multiple advanced machine learning models such as Logistic Regression, Random Forest, XGBoost, Voting classifiers.

2) To compare these models based on their accuracy, precision, recall, F1-score, ROC-AUC, and computational efficiency in predicting diabetic outcomes and classifying diabetic retinopathy risk.

3) To identify the most effective machine learning models for use in clinical settings, providing a foundation for targeted screening and personalized management strategies for patients at elevated risk of diabetic retinopathy.

Through these objectives, the study will contribute to the broader goal of reducing the incidence and impact of diabetic retinopathy by integrating sophisticated analytical techniques into the clinical decision-making process.

III. Materials and Methods

A. Data Description

The primary application of the Diabetes Prediction Dataset is in the development of predictive models using machine learning techniques. The dataset, sourced from Kaggle [22], is comprised of 100,000 electronic health records (EHRs) with nine features collected from multiple healthcare providers, used by researchers for research and analysis. It integrates medical and demographic data from patients diagnosed with or at risk of developing diabetes, emphasizing its utility for constructing machine learning models aimed at predicting diabetes likelihood. The dataset features include age, gender, body mass index (BMI), hypertension, heart disease, smoking history, hemoglobin A1c (HbA1c) levels, and blood glucose levels, each critical for assessing the patient's health status. Each entry is labeled with the diabetes status of the patient, categorized as positive or negative, allowing for the creation of machine learning models that can predict diabetes onset based on existing health data.

The dataset's demographic range includes precise age values, particularly for children under two years, represented in decimals (e.g., 0.08 equivalent to 1 month, 1.32 equivalent to 1 year and 4 months). This precision enables a nuanced understanding of diabetes risk factors across early age groups. For patients visiting Emergency Rooms, Hospitals, or Clinics, blood glucose levels were captured randomly, without specific fasting requirements, providing a broad but non-standardized snapshot of glucose regulation in potentially acute scenarios.

The dataset does not distinguish between type 1 and type 2 diabetes, making it crucial for predictive models to consider both types. Additionally, the smoking history variable categorizes individuals into six groups: never, not current, former, current, ever, and no info, reflecting varying degrees of exposure to smoking—a known risk factor for diabetes.

B. Data Collection Methodology

The data for this dataset was collected through various means including direct surveys, review of medical records, and laboratory tests from patients diagnosed with or at risk of developing diabetes. This approach ensures a comprehensive gathering of relevant health indicators which are critical in diabetes prediction. Post-collection, the data underwent rigorous processing to refine and standardize the information, ensuring its readiness for analytical applications [22].

The dataset includes 100,000 entries with demographic and medical attributes (0 for negative, 1 for positive).

The Diabetes Prediction dataset includes the following columns: gender: Gender of the patient. Three categories (Female, Male, Other), age: Age of the patient, hypertension: Whether the patient has hypertension (1) or not (0), heart disease: Whether the patient has heart disease (1) or not (0), smoking history: Smoking history of the patient. Six categories (No Info, current, ever, former, never, not current), BMI: Body Mass Index of the patient, HbA1c_level: Hemoglobin A1c level, a measure of average blood glucose over the past three months, blood glucose level: Current blood glucose level, and diabetes: Diabetes status (1 for positive, 0 for negative).

C. Exploratory Data Analysis

1) Observations: Age: Patients range from 0.08 to 80 years old, indicating inclusion of all age groups with an average age of approximately 41.89 years. Hypertension: 7.485% of patients have hypertension. Heart Disease: 3.942% of patients have heart disease. BMI: Ranges from 10.01 to 95.69 with a mean value of approximately 27.32, which indicates overweight on average according to the BMI scale. HbA1c level: Ranges from 3.5 to 9.0 with an average of 5.53, which is in the normal to slightly elevated range. Blood Glucose Level: Ranges from 80 to 300 mg/dL with a mean of approximately 138.06 mg/dL. Diabetes Status: 8.5% of the dataset is labeled as having diabetes mellitus.

2) Depth observation and analysis: In this section, an in-depth Exploratory Data Analysis (EDA) is conducted to understand the nuances of diabetes through several key objectives. Initially, the focus is on the distribution of crucial variables such as age, BMI, blood glucose levels, and HbA1c levels to establish baseline data behaviour (Fig. 1). The relationships these variables have with diabetes status are then explored, using advanced visualization techniques like pair plots. These plots specifically allow for the examination of interactions among the variables, categorized by diabetes mellitus status to discern patterns and anomalies effectively.

Further analysis leverages a robust Random Forest Machine Learning classifier to pinpoint the most significant predictors of diabetes. This model not only processes a vast amount of data but also provides insights into critical factors such as weight, sugar levels, age, and smoking history. Understanding these predictors aids healthcare professionals in early identification and intervention for those at high risk of developing diabetes mellitus.

The distribution plots in Fig. 1 effectively illustrate the key variables' distributions, highlighting the dataset's diversity and relevance for diabetes-related predictive analytics.

Age Distribution**:** The age distribution shows a relatively uniform spread across different age groups, with noticeable peaks in the younger and older populations. There is a significant increase in frequency around ages 70-80, indicating a higher number of elderly individuals in the dataset. The distribution suggests a broad age range, making the dataset suitable for age-related predictive analysis.

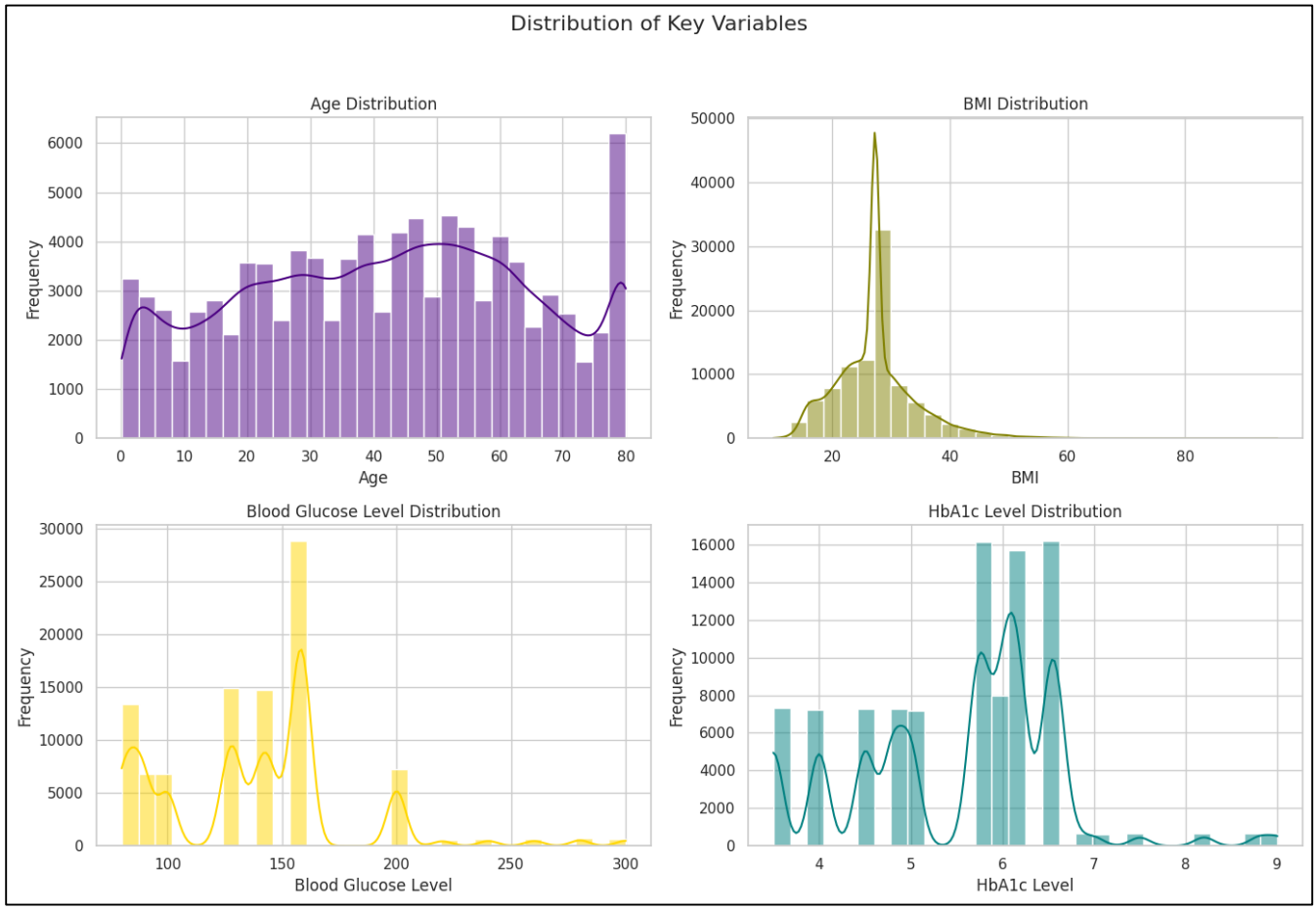


Fig. 1. Visual analysis of key variables.

BMI Distribution**:** The BMI distribution is skewed to the right, with most individuals having a BMI between 20 and 30. There is a noticeable peak around BMI 25, indicating that a significant portion of the population falls within the overweight category according to BMI classifications (BMI 25-29.9). This skewed distribution indicates a higher prevalence of overweight and moderately obese (BMI 30-34.9) individuals, which is relevant for diabetes and retinopathy risk assessment.

Blood Glucose Level Distribution**:** The distribution of blood glucose levels shows several peaks, with a significant one around 150 mg/dL. There are multiple smaller peaks indicating varying levels of blood glucose among the population. The distribution highlights a wide range of blood glucose levels, which is essential for predicting diabetes risk.

HbA1c Level Distribution**:** The HbA1c level distribution shows distinct peaks around values of 5, 6, and 7%. This indicates that there are clear clusters of individuals with specific HbA1c levels, which correspond to normal, pre-diabetic, and diabetic ranges. The presence of these clusters suggests that the dataset contains individuals across the spectrum of diabetes risk, from normal to high risk.

The relationships between these variables and diabetes mellitus status are then explored in Fig. 2 as follows:

- Diabetes Prevalence by Gender: In the female category, the count of non-diabetic individuals is significantly higher than that of diabetic individuals, with a noticeable but smaller group of diabetic females. For males, a similar pattern is observed: non-diabetic males have a higher count compared to diabetic males, though the number of non-diabetic males is less than non-diabetic females. The "Other" category has a very low count for both diabetic and non-diabetic individuals, indicating this category has fewer samples in the dataset.
- Diabetes Prevalence by Smoking History: Among those who have never smoked, most individuals are non-diabetic, but there is a small proportion of diabetics. The "No Info" category, similar to the "never" category, mostly consists of non-diabetic individuals, with a small number of diabetics. For current smokers, the count shows a higher number of non-diabetic individuals, but there is also a noticeable group of diabetics. In the former smokers category, there is a higher number of non-diabetic individuals, with a small number of diabetics. The "Ever" category, similar to "current" smokers, has more non-diabetic individuals with a small diabetic group. Finally, the "Not Current" category, which includes individuals who have smoked in the past but not currently, predominantly consists of non-diabetic individuals, with a smaller diabetic group.

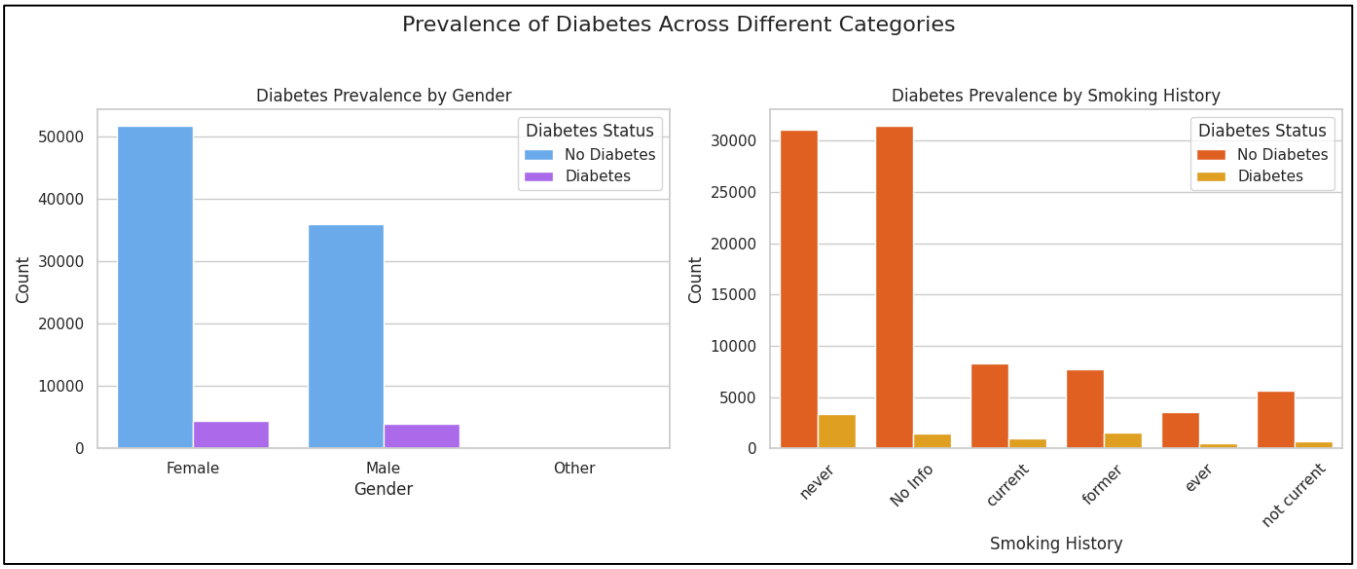


Fig. 2. Analysis of diabetes prevalence by gender and smoking history.

Factor Analysis was done where advanced visualizations were employed to explore the relationships between key variables related to diabetes mellitus (Fig. 3). Specifically, pair plots are utilized to analyze the interactions among age, BMI, blood glucose levels, and HbA1c levels. These plots segment data by diabetes status (0 for non-diabetic and 1 for diabetic), enabling to identify patterns and outliers clearly. This visualization helps highlight how these variables correlate with each other and their collective impact on diabetes prevalence.

Fig. 3 provides a detailed interpretation of the visual data gathered:

- In the Age and Diabetes: There is a noticeable density peak in the age distribution among diabetics at around 55-70 years, indicating a higher prevalence of diabetes in this age group compared to younger individuals.
- BMI: The distribution of BMI values is similar across both diabetics and non-diabetics. However, there is a slightly higher density of diabetic individuals with a BMI above 30, suggesting a potential link between higher BMI and increased diabetes prevalence.
- Blood Glucose Level: Generally, diabetic individuals display elevated blood glucose levels, as evidenced by the clustering of green dots (diabetics) above typical threshold values.
- HbA1c Level: There is a clear distinction in HbA1c levels, with diabetic individuals typically showing higher levels, often exceeding 6.5%, a commonly used diagnostic threshold for diabetes.
- Inter-variable Relationships: The data reveals notable patterns, such as the positive relationship between BMI and blood glucose level, as well as between BMI and HbA1c level, which are more pronounced in diabetic individuals.

Coding Reference: For detailed technical insights, please refer to Appendix A, which contains the GitHub repository link.

Overall, the visualization indicates distinct distributions for diabetic individuals in terms of blood glucose and HbA1c levels and suggests a correlation between age, BMI, and the likelihood of having diabetes. The relationships presented can inform healthcare professionals in identifying high-risk profiles and tailoring interventions accordingly.

3) Key factors for predicting diabetes: The study employs a robust Random Forest machine learning classifier to identify the factors that most significantly affect the likelihood of developing diabetes. A comprehensive dataset is analyzed using the Random Forest algorithm, which serves as a powerful tool to determine key predictors of diabetes. This analysis highlights important indicators such as weight, sugar levels, age, and smoking habits. This helps doctors figure out who might get diabetes and how to help them early (Fig. 4).

The bar plot in Fig. 4 visualizes the feature importance determined by a Random Forest classifier for predicting diabetes. HbA1c level and blood glucose level are the top factors, indicating their strong predictive power for diabetes. BMI and age are also significant, whereas smoking history, hypertension, heart disease, and gender have less influence on the model's predictions.

The Random-Forest classifier has provided the following feature importance, which indicates how much each feature contributes to the model's ability to predict diabetes (Fig. 4):

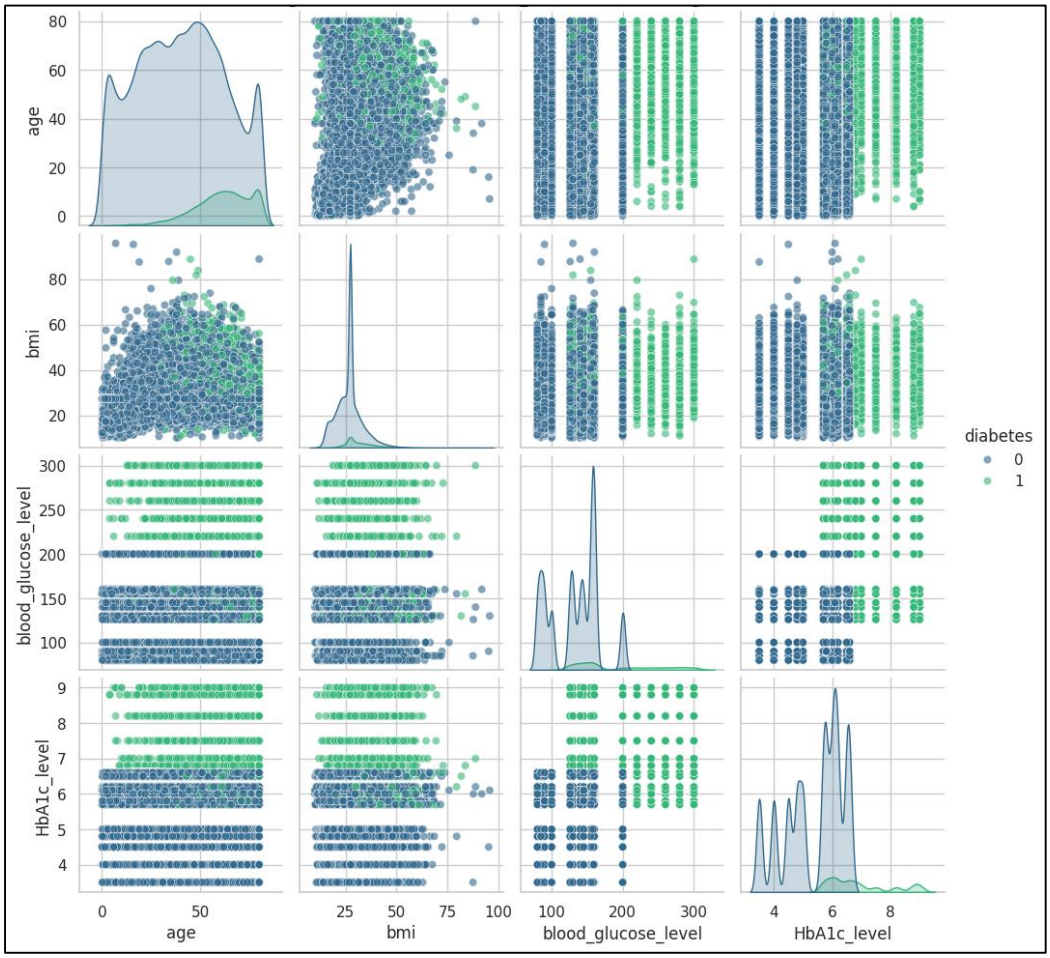


Fig. 3. Pair plot of key variables segmented by diabetes status.

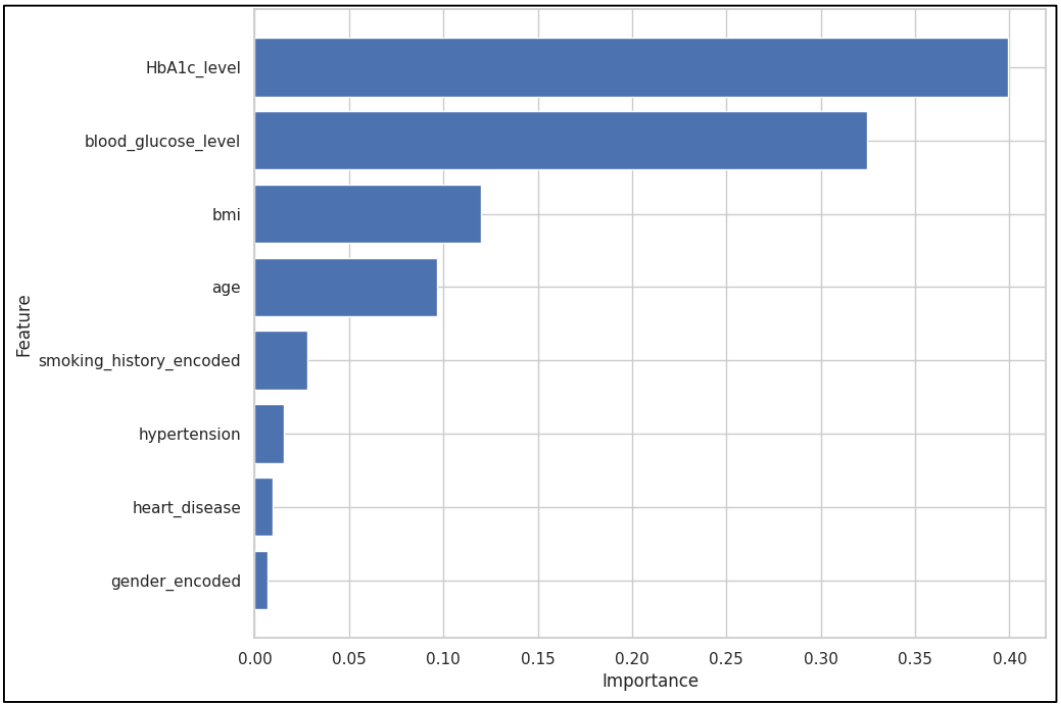


Fig. 4. Feature importance for diabetes prediction.

HbA1c_level (39.31%): The most important feature. HbA1c levels reflect average blood glucose levels over the past few months, making it a critical indicator of diabetes. Blood Glucose Level (32.67%): The second most significant predictor, which is directly related to diabetes, as it measures the current sugar levels in the blood. BMI (12.06%): Body Mass Index also plays a significant role, reflecting the obesity level which is a known risk factor for diabetes. Age (9.88%): Age is another important factor, as the risk of developing diabetes increases with age. Smoking History Encoded (2.73%): Smoking history has a moderate impact, potentially due to its influence on general health and cardiovascular risk, which is related to diabetes. Hypertension (1.59%): Hypertension is moderately important, likely due to its association with cardiovascular health. Heart Disease (1.03%): Similarly, heart disease shows a small impact, which correlates with overall metabolic health. Gender Encoded (0.72%): Gender has the least importance according to this model, suggesting it has a minimal direct impact on diabetes risk in this dataset.

These results help identify which features are most predictive of diabetes in the dataset and can guide further data analysis, feature engineering, and the development of intervention strategies.

D. Data Pre-processing

Data preprocessing is a critical step that involves preparing the raw data for machine learning models. This step typically involves several sub-steps:

1) Cleaning: Includes handling missing values and removing duplicates. This is crucial because missing values can introduce bias or inaccuracies into the model, and duplicates can lead to overfitting and the result skew the model training.

$X_{cleaned}$=f(dropna,drop_duplicates(X)), Where X represents the initial dataset.

Equation: Cleaned Data=Raw Data − (Missing Values + Duplicates)

In this context, the equation represents the removal of undesirable data elements, ensuring that only valid, unique data points are used for further analysis.

Data Cleaning Summary

- Missing Values: There are no missing values in any of the columns.
- Duplicate Rows: There are 3,854 duplicate entries in the dataset.
- Data Consistency: There are no negative values in columns such as 'age', 'bmi', 'HbA1c_level', or 'blood glucose level'.

2) Encoding categorical variables: The dataset includes a mix of categorical and numerical variables:

- Categorical: gender, hypertension, heart disease, smoking history, diabetes (target variable to be used for prediction).
- Numerical: age, bmi, HbA1c_level, blood glucose level

Since many machine learning models require a format suitable for machine learning models which is numerical input. Therefore, categorical variables need to be converted into a numerical format. One-hot encoding is a common technique used where each categorical value is converted into a new categorical column and assigned a 1 or 0.

$$X_{encoded}=\text{OneHotEncoder}(X_{categorical})$$

3) Correlation analysis: The relationships among these variables were explored using a correlation matrix. This will help identify which factors are most strongly associated with diabetes. As shown in Fig. 5, the correlation matrix highlights relationships between features: Age shows a mild positive correlation with diabetes, indicating that risk increases with age. Hypertension and heart disease also show positive correlations with diabetes status, suggesting that these conditions are associated with higher diabetes risk. BMI has a slight positive correlation with diabetes, supporting the known link between obesity and increased diabetes risk. The blood glucose level and HbA1c level have stronger positive correlations with diabetes, as expected, since they directly measure aspects of blood sugar management.

E. Model Design

Fig. 6 illustrates a comprehensive workflow for predicting diabetes and categorizing the risk of retinopathy. It is divided into two main phases: Phase I involves the development and evaluation of machine learning models for diabetes prediction, while Phase II focuses on assessing the risk of retinopathy for patients identified as diabetic. This systematic approach ensures accurate prediction and effective risk stratification, facilitating timely and appropriate medical interventions.

In the same context, Fig. 7 depicts a visual representation of the entire workflow for the diabetes prediction system. It outlines each step from the initial data acquisition to the final model evaluation. The process begins with the collection and cleaning of the diabetes dataset, followed by various preprocessing techniques to prepare the data for machine learning algorithms [30]. It includes steps such as data scaling, encoding, and addressing class imbalances. The diagram also illustrates the model selection, hyperparameter tuning, and cross-validation processes, culminating in the deployment of the most effective model for diabetes prediction. This systematic approach ensures the development of a robust and reliable prediction system (Fig. 6 and Fig. 7).

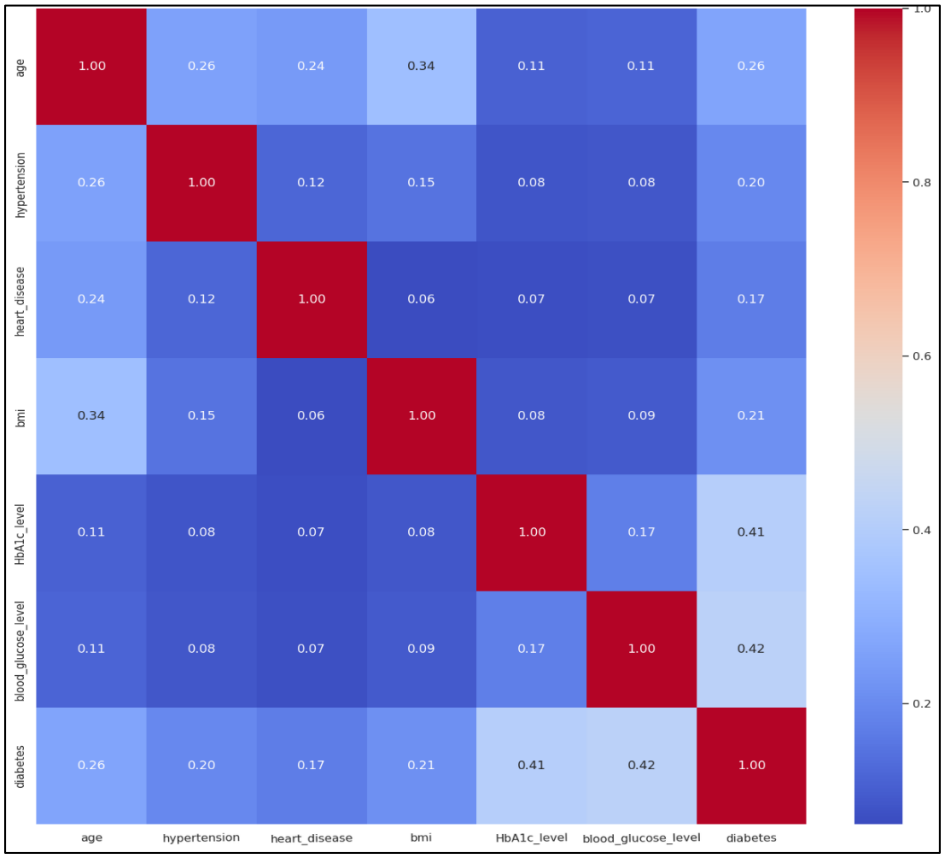


Fig. 5. Features' relationships correlation matrix.

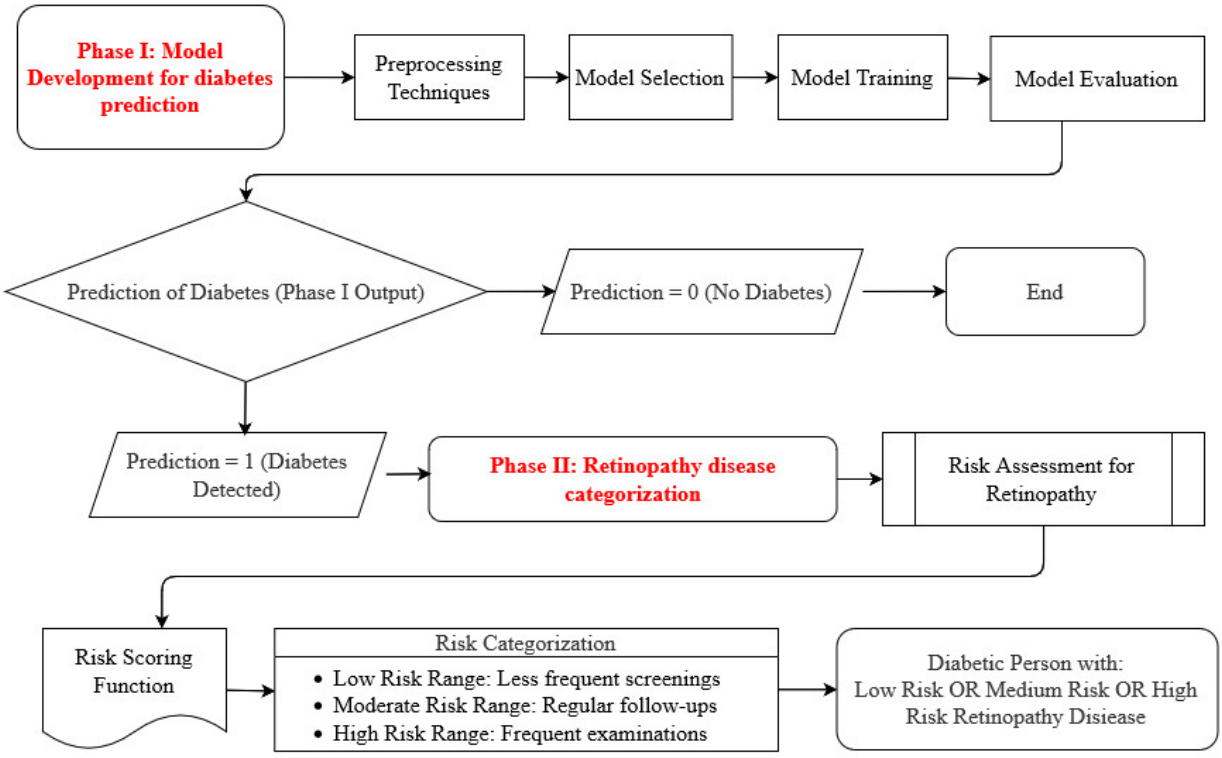


Fig. 6. Flowchart: steps of conducting the study.

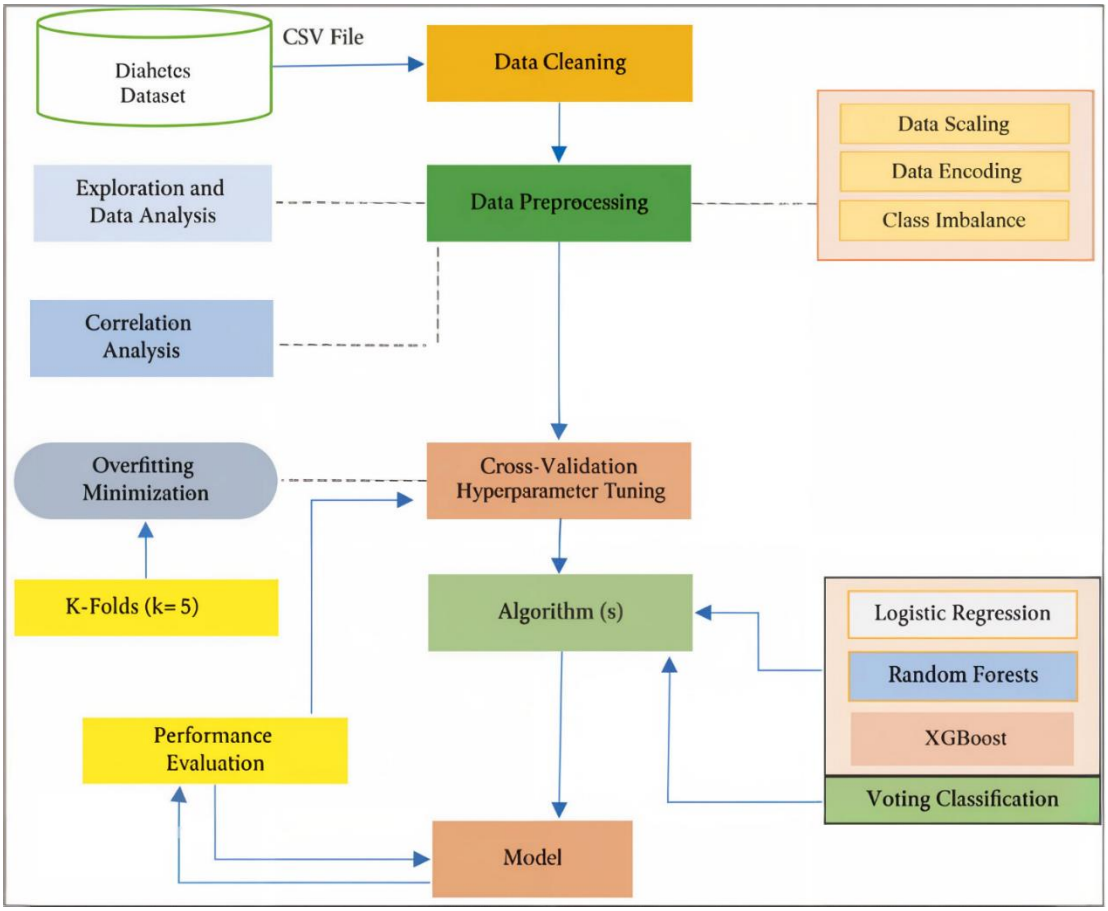


Fig. 7. Block diagram and operational mechanism flow for machine learning in diabetes prediction.

1) Phase 1 - machine learning for diabetes predictions: In the first phase of the study, the objective was to develop and fine-tune machine learning models capable of predicting diabetes. This involved the following steps:

a) Model selection: In selecting models for diabetes predictions, algorithms were chosen for their ability to effectively handle the complexity and variability of medical data, ensuring both high predictive accuracy and robustness against overfitting. This strategic selection helps tailor the approach to accurately capture the nuanced relationships within diabetes-related variables. In predicting diabetes, the selection of machine learning models is critical due to the need for high accuracy and the ability to generalize well from medical datasets. Here's a deeper look into the significance and roles of each chosen model:

- Logistic Regression: This model serves as a fundamental baseline in medical prediction tasks due to its simplicity and interpretability. It uses a logistic function to estimate probabilities, making it particularly useful for binary outcomes like diagnosing diabetes. Its coefficients provide insights into the influence of each feature, aiding clinicians in understanding risk factors.

Objective function: $\hat{y} = \sigma(x\beta + b)$, $\sigma(z) = \frac{1}{1-e^{-z}}$ where σ is the logistic function, β is the coefficient vector, b is the bias, and $\hat{y}$ is the predicted probability [23].

- Random Forest: As an ensemble of decision trees, Random Forest mitigates the risk of overfitting associated with individual decision trees by averaging multiple predictions, thereby enhancing the model's stability and accuracy. Its ability to handle large datasets with many features makes it invaluable for capturing complex, nonlinear relationships that are typical in medical data.

Objective Function: $\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i \cdot (x)$, Where T_i represents an individual tree's prediction and N is the number of trees [24].

- XGBoost: Known for its efficiency and performance, XGBoost is a sophisticated version of gradient boosting that has proven to be extremely effective in various Kaggle competitions involving medical predictions. It optimizes both speed and prediction accuracy by building trees sequentially, each one correcting errors made by the previous, which is crucial for a nuanced disease like diabetes where early detection can significantly alter patient outcomes.

Objective Function: $\hat{y}_i = \sum_{k=1}^{k} f_k \cdot (x_i)$, where f_k € F. F is the space of trees and f_k represents an individual tree [24].

- Voting classifier: This ensemble technique combines predictions from the Logistic Regression, Random

Forest, and XGBoost [25] models. By using a soft voting mechanism, it computes the final output based on the probability estimates from each model, rather than simple majority rules. This approach helps in reducing variance and bias, leveraging the strengths while balancing the weaknesses of the constituent models, resulting in more reliable and robust prediction capabilities [26].

These models are selected not only for their individual merits but also for their collective ability to provide a comprehensive understanding of the predictive landscape. This ensemble strategy enhances predictive performance, ensuring that the diagnostic tool is both accurate and reliable in a clinical setting.

b) Preprocessing Techniques [27][28][29]: Preprocessing techniques in ML involve cleaning and transforming raw data to improve model performance. Common methods include handling missing values, normalizing data, encoding categorical variables, and addressing class imbalances. Fig. 8 illustrates the operational mechanism of ML models designed for predicting diabetes.

The effectiveness of machine learning models significantly depends on the quality of the data they are trained on. Therefore, rigorous data preprocessing is imperative. Hence, the data was further processed using the following techniques to enhance predictions:

- Data Scaling [31]**:** All numeric features were scaled using the Standard Scaler method to normalize the distribution, aiding in faster convergence during the training phase. Data is standardized to have zero mean and unit variance. Xscaled $=\frac{x-\mu}{\sigma}$, where μ and σ are the mean and standard deviation of the features, respectively. Standardization is crucial for models that are sensitive to the scale of input data.

- Handling Class Imbalance [31]**:** The SMOTE (Synthetic Minority Over-sampling Technique) algorithm was employed to address class imbalance in the dataset, ensuring that the minority class is adequately represented during model training. To address class imbalance in the dataset, SMOTE is applied: $X_{resampled}$, $y_{resampled}$ = SMOTE(X_{train}, y_{train}). SMOTE generates synthetic samples from the minority class, making the class distribution equal and thus preventing model bias towards the majority class.

- Cross-Validation [32]: Stratified K-Fold cross-validation with five folds was employed, which is particularly useful for imbalanced datasets. This method ensures that each fold of the dataset has the same proportion of examples in each class as the complete set. This approach provides a robust estimate of the model's performance, as it iteratively trains the model on k−1 folds and validates it on the remaining fold, cycling through all k folds as the validation set. It provides confidence that the models are stable and perform well across different subsets of the dataset, reducing the likelihood of model overfitting and ensuring that the predictions are reliable.

- Hyperparameter Tuning [33]: GridSearchCV was implemented to automate the process of tuning parameters to find the best combination for each model. This exhaustive search over specified parameter values for an estimator is crucial for optimizing the learning algorithm. Each model was assessed using the ROC-AUC score as the scoring metric, which measures the ability of the model to distinguish between the classes across all possible thresholds. This tuning of parameters optimizes model performance on the dataset, ensuring that the predictions are as accurate as possible, which is critical for medical applications where the cost of false predictions can be high.

- Optimal Parameters and Model Evaluation: After tuning, the optimal parameters for each model were established and used to train the models on the processed training set. The evaluation of these models on a hold-out test set involved the following metrics:

```
- 'Logistic Regression': LogisticRegression(C=0.01)
- 'Random Forest': RandomForestClassifier(max_depth=20,
  n_estimators=200)
- 'XGBoost': XGBClassifier(learning_rate=0.1,
  max_depth=6, n_estimators=150)
- 'Voting Classifier': VotingClassifier( estimators=[ ('lr',
  LogisticRegression(C=0.01)), ('rf',
  RandomForestClassifier(max_depth=20,
  n_estimators=200)), ('xgb',
  XGBClassifier(learning_rate=0.1, max_depth=6,
  n_estimators=150)) ], voting='soft'
```

These parameters were then used to train each model on the entire training set processed through SMOTE and scaled appropriately. The trained models were evaluated on a hold-out test set to gauge their effectiveness, using metrics such as accuracy, precision, recall, F1-score, and the ROC-AUC.

- Model Training**:** After preprocessing the data (as detailed in previous discussions), four key models were trained using the best hyperparameters identified through GridSearchCV. These models included Logistic Regression, Random Forest, XGBoost, and Voting.

- Model Evaluation**:** Each model was rigorously evaluated on a split test set to assess their performance in accurately predicting diabetes. Metrics such as accuracy, precision, recall, F1-score, and AUC-ROC were employed to compare each model's effectiveness.

c) Performance Evaluation [34][35]: To evaluate the effectiveness of different models, several metrics are used:

- Accuracy**:** The proportion of true results among the total number of cases examined.

- Precision**:** The proportion of positive identifications that were actually correct.

- Recall**:** The proportion of actual positives that were correctly identified.

- F1-Score**:** The harmonic mean of precision and recall.

- ROC-AUC: The area under the receiver operating characteristic curve, which plots the true positive rate against the false positive rate at various threshold settings.

Equations:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = 2\times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Where TP represents instances where the model correctly predicts a positive outcome. FP refers to cases where the model predicts a positive outcome, but the actual result is negative. TN indicates instances where the model correctly predicts a negative outcome, while FN refers to cases where the model predicts a negative outcome, but the actual result is positive.

- Running Time: Running time indicates the computational efficiency of each model.
- Confusion Matrix: The confusion matrix provides a detailed breakdown of true positive, false positive, true negative, and false negative predictions. It compares the actual target values with those predicted by the machine learning model, providing a holistic view of the model's performance and highlighting the types of errors it makes (Fig. 8).

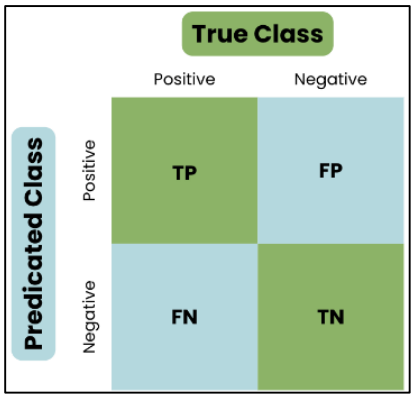


Fig. 8. The basic structure of a confusion matrix.

d) Selection of optimal models: Based on the evaluation, models that exhibited the highest efficacy in terms of AUC-ROC and F1-score are favored. This ensured that the chosen models were not only accurate but also balanced in terms of precision and recall.

e) Significance of accurate diabetes predictions: Accurate diabetes predictions enable timely interventions that can prevent or delay the onset of complications like retinopathy. By identifying individuals at risk of developing diabetes or managing those already diagnosed more effectively, healthcare providers can implement preventative measures such as lifestyle modifications, regular monitoring, and early pharmacological intervention.

2) Phase 2 - Application on new dataset and risk assessment for retinopathy: Upon successfully training and selecting the best models, the next phase involved applying these models to a new dataset (Appendix B). This dataset comprised unseen data, simulating a real-world scenario where the models predict diabetes status in new patients. Following the prediction, a detailed risk assessment for diabetic retinopathy was conducted:

- Prediction on New Data: The trained models were applied to the new dataset to predict diabetes (Appendix B). This step tested the models' generalizability and their ability to function accurately outside the training data environment.
- Risk Assessment for Retinopathy: Risk Scoring Function: A custom function was developed to assign scores to various features based on their significance and impact on retinopathy risk. This included typical and atypical values and ranges for features such as HbA1c levels, hypertension, BMI, age, and smoking history. Each feature's contribution to the risk score was weighted according to established medical research indicating its influence on retinopathy.

Table I summarizes the scoring rules and conditions that represent how different factors contribute to a risk score when predicting diabetes. It provides a comprehensive scoring system to predict diabetes risk by evaluating various health and lifestyle factors. Each condition is assigned a score based on specific ranges or values, contributing to an overall risk score. The factors include HbA1c level, hypertension, heart disease, BMI, age, smoking history, gender, and blood glucose level. By aggregating the scores from these conditions, healthcare providers can better classify patients' risk levels and prioritize interventions.

- Retinopathy Risk Categorization (only for diabetic predicted patients): Based on the cumulative risk score derived from the scoring function, each patient was categorized into No, Low, Medium, or High Risk for developing diabetic retinopathy.
  - No Risk (score range 0-5): Diabetic patients with no risk of retinopathy do not require additional retinal screenings.
  - Low Risk (score range 5-7): Patients predicted with low probability of diabetes might require less frequent retinal screenings.
  - Moderate Risk (score range 7-9): Patients showing borderline or moderate probabilities may need more regular follow-ups to monitor any progression in retinal changes.
  - High Risk (score range >9): Patients predicted to be highly likely to have or develop diabetes should undergo comprehensive and possibly more frequent retinal examinations to detect early signs of retinopathy.

TABLE I. RETINOPATHY SCORING CONDITIONS TABLE

Condition	Range/Value	Score
Predicted Model Output	0 – Non-Diabetic	0
	1 – Diabetic	Continue scoring
HbA1c Level	≤ 7	0
	> 7 and ≤ 8	1
	> 8 and ≤ 9	2
	> 9	3
Hypertension	0	0
	1	1
Heart Disease	0	0
	1	1
BMI	< 25	0
	≥ 25 and < 30	1
	≥ 30	2
Age	< 40	0
	≥ 40 and < 50	1
	≥ 50 and < 60	2
	≥ 60	3
Smoking History	'never', 'No Info'	0
	'ever', 'not current'	1
	'former'	2
	'current'	3
Gender	'Other'	0
	'Male'	0
	'Female'	0
Blood Glucose Level	< 100	0
	≥ 100 and < 126	1
	≥ 126 and < 200	2
	≥ 200	3

This categorization helps in prioritizing medical attention and preventive measures.

IV. RESULTS

Predictions were conducted on testing data, comprising 20% of the original dataset in the first phase, and on a completely new dataset (Appendix B) in the second phase. Predictions are structured such that a result of 0 indicates the patient does not have diabetes, thereby assigning a risk score of zero for retinopathy. Conversely, if the prediction indicates diabetes, the patient's clinical results must undergo a risk assessment scoring system. This system categorizes patients as low, medium, or high risk, based on the severity of their scores. In the second phase, machine learning models are trained on the entire original dataset to leverage known diabetes outcomes and apply predictions to a new, separate dataset. This approach enables the prediction of diabetic status and subsequent assessment and categorization of retinopathy risk.

A. *Phase 1: Classifiers' Result*

In Phase 1 of the study, the performance of various classifiers on the task of predicting diabetes was evaluated. The classifiers tested include Logistic Regression, Random Forest, XGBoost, and a Voting Classifier. Each model was assessed based on several performance metrics: Accuracy, Precision, Recall, F1-Score, AUC-ROC, Running Time, and Confusion Matrix. The results, summarized in Table II, provide a comprehensive comparison of the models' effectiveness and efficiency in diabetes prediction.

TABLE II. PERFORMANCE METRICS OF CLASSIFIERS IN DIABETES PREDICTION

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)	Confusion Matrix
Logistic Regression	0.87	0.41	0.86	0.55	0.95	4.9359	[[15461, 2073], [237, 1459]]
Random Forest	0.94	0.66	0.78	0.72	0.96	882.9681	[[16871, 663], [364, 1332]]
XGBoost	0.96	0.84	0.73	0.78	0.97	45.2544	[[17300, 234], [451, 1245]]
Voting Classifier	0.95	0.73	0.77	0.75	0.96	15.0121	[[17071, 463], [385, 1311]]

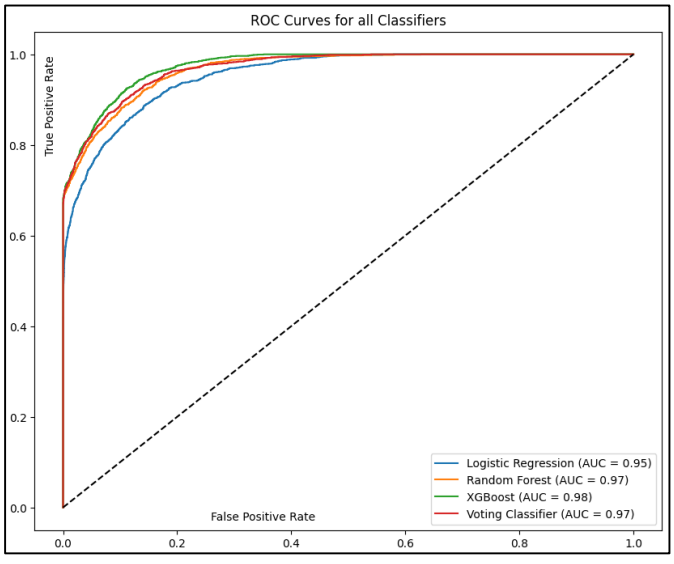


Fig. 9. ROC curves for all classifiers.

Additionally, the ROC Curves for all classifiers, depicted in Fig. 9, illustrate the true positive rate versus the false positive rate, providing insight into the models' ability to distinguish between classes. This visual representation, combined with the detailed performance metrics, helps to identify the strengths and weaknesses of each classifier, guiding the selection of the most suitable model for further development.

1) Comparative analysis of classifiers for diabetes prediction: In the realm of diabetes prediction, the performance of classifiers can significantly influence the effectiveness of diagnosis and subsequent patient management. The evaluation of four classifiers—Logistic Regression, Random Forest, XGBoost, and a Voting Classifier—provides insights into their efficacy across various metrics that are crucial for medical decision-making.

Logistic Regression is notable for its high recall of 86.03%, indicating its ability to identify a high number of true positive cases, which is critical in medical diagnostics to ensure that few cases of diabetes go undetected. However, its precision is relatively low at 41.31%, suggesting a higher rate of false positives that could lead to unnecessary anxiety or treatment. Despite these trade-offs, its rapid processing time of under 5 seconds and an AUC-ROC score of 95.48% demonstrate its utility in scenarios where speed and general accuracy are prioritized over precision.

Random Forest shows a marked improvement in overall accuracy (94.66%) and precision (66.77%) compared to Logistic Regression, suggesting better balance in identifying true positives while reducing false positives. Its recall of 78.54% remains robust, albeit lower than Logistic Regression, reflecting a more conservative but precise approach. The main limitation of Random Forest in this context is its computational demand, with a significantly longer running time, which might be a constraint in real-time prediction environments.

XGBoost emerges as the strongest performer in terms of accuracy (96.44%) and AUC-ROC (97.54%), underscoring its capability to effectively separate the diabetic and non-diabetic classes. With the highest precision (84.18%) among the classifiers, XGBoost offers a reliable prediction model that minimizes false positives—a desirable feature in clinical settings. Nevertheless, the trade-off here involves its recall (73.41%), which is lower than Logistic Regression's, pointing towards a potential underdiagnosis risk.

The Voting Classifier combines the strengths of the above models and achieves an accuracy of 95.59%, with well-balanced precision (73.90%) and recall (77.30%). This classifier harnesses the collective insights of Logistic Regression, Random Forest, and XGBoost, potentially leading to more consistent predictions across diverse patient profiles. The Voting Classifier's middle-range running time (15.01 seconds) and high AUC-ROC (96.95%) make it a viable option for both accuracy and efficiency in clinical applications.

2) Further experiments: This section explores machine learning techniques for improving diabetes prediction accuracy by addressing class imbalance and optimizing prediction thresholds. Detailed quantitative results are provided in Appendices C and D.

a) Handling Class Imbalance (Appendix C): First, SMOTE (Synthetic Minority Over-sampling Technique) was used to address class imbalance, which significantly improved recall and minimized false negatives in diabetes prediction (Table A1). ADASYN (Adaptive Synthetic Sampling) was also tested, focusing on generating samples near decision boundaries, but it did not surpass SMOTE's performance (Table A2). Additionally, experiments with BorderlineSMOTE (Table A3), which selectively generates samples around the decision boundary, yielded mixed results, confirming SMOTE as the primary method.

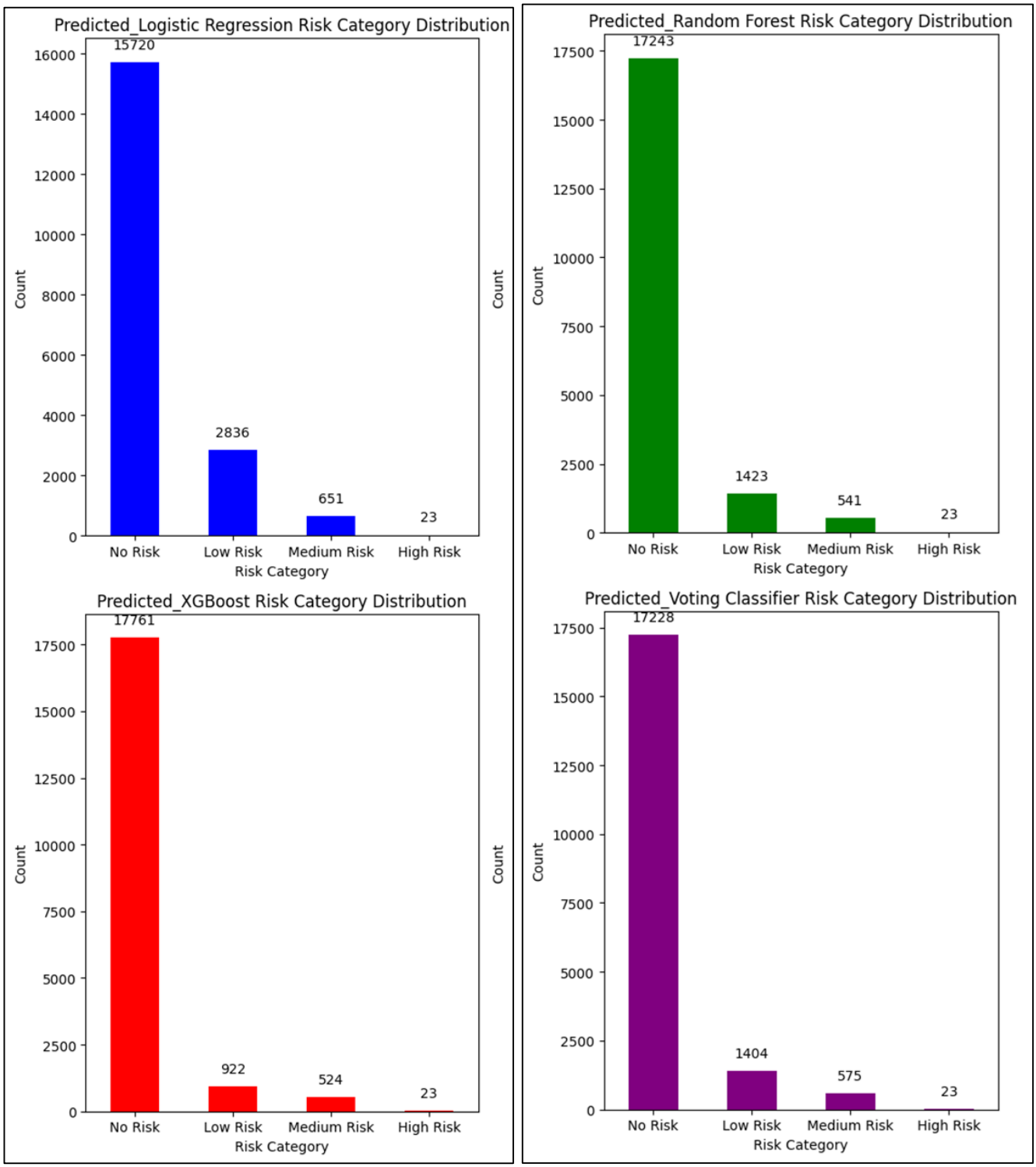
b) Optimizing Prediction Thresholds (Appendix D): To enhance clinical utility, prediction thresholds were adjusted to improve recall, aiming to reduce false negatives. Thresholds of 0.5, 0.6, 0.25, and 0.4 were tested, observing their impacts on precision, recall, and F1-score. A lower threshold (0.25 or 0.4) maximized recall, suitable for screening to identify as many positive cases as possible. A higher threshold (0.6) improved precision, suitable for diagnostic settings where false positives are costly. Detailed metrics for these adjustments are in Tables B1 to B3 (Appendix D).

These experiments highlight the importance of tailored approaches in machine learning for healthcare. Adjusting class imbalance handling and prediction thresholds can significantly enhance model performance and suitability for specific healthcare applications, particularly in early and accurate diabetes detection.

B. Phase 2: Retinopathy Risk Assessments' Result

In Phase 2 of the study, the risk of retinopathy was assessed using various classifiers. The classifiers tested include Logistic Regression, Random Forest, XGBoost, and a Voting Classifier. The goal is to categorize individuals into different risk levels: No Risk, Low Risk, Medium Risk, and High Risk.

The distribution of predicted risk categories for each classifier is illustrated in Fig. 10A and Fig. 10B. All models heavily favor "No Risk" predictions, indicating an imbalanced dataset with more non-risk instances. Very few predictions in the "High Risk" category across all models. Logistic Regression and Random Forest are more balanced compared to the conservative XGBoost, which shows the highest "No Risk" predictions. The Voting Classifier balances predictions, indicating the benefit of ensemble methods for nuanced risk detection. These insights can guide model selection with balanced parameters and dataset management for more accurate and trustworthy AI predictions.



(a) (b)

Fig. 10. (a) A Classifiers risk category distribution_XGBoost, (b) Classifiers risk category distribution_Voting Classifier

V. Discussion and Conclusion

The comparative analysis reveals that while XGBoost offers the highest precision and overall accuracy, making it suitable for settings where the cost of a false positive is high, the Voting Classifier provides a balanced solution that might be preferred in clinical environments where both types of errors (false positives and false negatives) carry significant consequences. Logistic Regression, with its high recall, could be particularly useful in initial screening tests were missing a positive diagnosis could be detrimental. Random Forest, with its strong performance across metrics but slower execution, might be more applicable in situations where computational time is less of a

constraint. The results demonstrate significant improvements in predictive accuracy compared to traditional models. For instance, the XGBoost model achieved an accuracy of 96.43%, which is approximately 9.60% higher than the commonly used logistic regression model, 1.88% higher than the random forest model, and 0.88% higher than the voting classifier. Additionally, the study identified novel patterns and risk factors that were previously unreported in the literature. These findings address critical gaps in existing research, particularly in the early detection and risk assessment of diabetic retinopathy.

The present study's use of XGBoost and Random Forest models for diabetes prediction demonstrated accuracies of 96.43% and 94.65%, respectively. These results, although slightly lower than the 97.82% accuracy reported for Random Forest by Alam et al. (2024) [36], reflect the comprehensive preprocessing steps undertaken, which were not fully implemented in the referenced studies, thereby enhancing the reliability and robustness of the findings. Furthermore, the Logistic Regression model in this study achieved an accuracy of 87.98%, compared to 96.06% reported by Gaur et al. (2024) [36]. Additionally, Voting Classifier achieved an accuracy of 95.59%, and KNN achieved 95.28%, aligning closely with the findings of Gaur et al. (2024) who reported 96.02% for KNN and 96.45% for SVM. Alshenawy and Almetwally (2023) [37] reported the highest accuracy for KNN at 99.99%, which underscores the potential of advanced models. Notably, our study also evaluated running time, revealing that XGBoost (45.25 seconds) and Voting Classifier (15.01 seconds) were more efficient than Random Forest (882.97 seconds), an aspect not considered in previous studies. This highlights the practicality of the models in real-life scenarios, where computational efficiency is crucial.

The choice of classifier in diabetes predictions should align with specific clinical priorities—whether it is reducing the risk of undiagnosed cases or minimizing the burden of false positives on the healthcare system. Each classifier has its strengths and scenarios where it might perform optimally, emphasizing the importance of context in model selection for medical applications.

When it comes to assessing the risk of diabetic retinopathy, the choice of classifier for diabetes and its complication predictions involves balancing various factors including accuracy, speed, and the specific medical consequences of false positives and false negatives. High-performing classifiers that effectively balance precision and recall, such as XGBoost and the Voting Classifier, are particularly valuable in these settings. Their use helps in creating stratified medical responses that optimize care for each patient based on their individualized risk profile, potentially leading to better clinical outcomes and more efficient use of healthcare resources.

Classifiers can categorize patients based on the likelihood of disease progression. The performance of each classifier can impact the assessment. Models with higher precision, such as XGBoost in this analysis, are crucial in this context. High precision reduces false positives, which means fewer patients are incorrectly categorized as at high risk of retinopathy. This is vital to avoid unnecessary treatments, which can be invasive and costly.

High recall rates are equally important because they ensure that most patients who are at risk of retinopathy are correctly identified for further testing and early treatment. Logistic Regression showed the highest recall, suggesting it could be useful in initial screening phases to ensure comprehensive identification of at-risk individuals. High overall accuracy and AUC-ROC, as seen with XGBoost and the Voting Classifier, indicate strong overall performance in distinguishing between patients at different levels of risk. This is essential for categorizing patients accurately into risk groups, which can guide the intensity and frequency of monitoring and intervention. In environments where real-time analysis is critical—such as in clinical settings during patient visits—models with shorter running times like Logistic Regression may be preferable despite other limitations.

The analysis underscores the crucial role of sophisticated machine learning classifiers in enhancing diabetes management and preventing its complications, notably diabetic retinopathy. Accurate diabetes prediction models can lead to early identification of individuals at risk, facilitating timely interventions that can significantly mitigate the progression of the disease and its associated complications.

The comparative analysis of different classifiers such as Logistic Regression, Random Forest, XGBoost, and Voting Classifier reveals that no single model fits all scenarios. Each classifier brings its strengths in terms of precision, recall, accuracy, and operational efficiency. For instance, XGBoost stands out for its high precision and accuracy, making it particularly useful in settings where reducing false positives is crucial. Meanwhile, Logistic Regression, with its high recall, is invaluable for initial screenings to ensure comprehensive identification of potentially at-risk individuals.

The choice of a classifier can significantly impact clinical outcomes. Precision in predictions minimizes the risk of unnecessary treatments, which is particularly important in managing diabetic retinopathy, where interventions can be as severe as laser surgery or injections. High recall is essential to avoid missing any cases of potential diabetes and its complications, ensuring that all at-risk individuals are monitored and treated appropriately.

The integration of these classifiers into healthcare systems implies a move towards more personalized medicine. It enables healthcare providers to categorize patients not just based on static factors but also through dynamic, data-driven insights, allowing for tailored monitoring schedules and treatments. This approach not only improves patient outcomes but also optimizes resource allocation within healthcare systems.

Unlike previous studies [36] [37] [38] that primarily focused on predicting diabetes alone, this research extends to evaluating the risk of diabetic retinopathy based on available data. If more features and detailed data were available, it could potentially extend to other diabetes-related complications. This study's approach of integrating multiple machine learning techniques, comparing them in terms of various metrics including computational efficiency, and analyzing a comprehensive dataset provides a more robust and accurate prediction framework. Novel risk factors were identified that were not highlighted in previous studies, addressing critical gaps in

existing research. The study also ensures that all necessary preprocessing steps are implemented, enhancing the reliability and robustness of the findings by making the data well-prepared for machine learning applications without introducing bias. Additionally, the study uniquely evaluates the running time of each model, highlighting practical efficiency and applicability in real-life scenarios. This aspect was not fully addressed in previous studies. By categorizing patients into preliminary risk levels for retinopathy, the work helps reduce the cost of unnecessary eye scans and other related examinations. The improved predictive accuracy enables earlier detection and intervention for at-risk patients, potentially reducing the incidence of severe complications and associated healthcare costs. This research provides a scalable and effective tool for diabetes and retinopathy risk evaluation, contributing significantly to the field by offering broader practical implications for healthcare providers.

Thus, this study proposes an automatic diabetes prediction system that can be deployed on a website and an Android smartphone application using the XGBoost machine learning framework. Users can input relevant data such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level. The system will provide instantaneous diabetes prediction along with the risk of retinopathy through the designed web application using real data.

There is a clear need for ongoing research and development in this area to refine these models, reduce their computational demands, and enhance their adaptability to real-world clinical settings. Additionally, the adoption of these technologies must be accompanied by training for healthcare professionals to maximize the benefits of such advanced tools.

Ultimately, leveraging advanced classifiers for diabetes prediction and retinopathy risk assessment represents a significant step forward in the fight against diabetes and its debilitating complications. As technology advances, the potential for these tools to become integral components of personalized healthcare grows, promising not only better patient outcomes but also more efficient healthcare systems globally.

REFERENCES

[1] World Health Organization. Global report on diabetes. World Health Organization; 2016.

[2] Desai, S., & Deshmukh, A. (2020). Mapping of Type 1 Diabetes Mellitus. Current diabetes reviews, 16(5), 438–441. https://doi.org/10.2174/1573399815666191004112647

[3] Hemmingsen B, Gimenez-Perez G, Mauricio D, Roqué i Figuls M, Metzendorf MI, Richter B. Diet, physical activity or both for prevention or delay of type 2 diabetes mellitus and its associated complications in people at increased risk of developing type 2 diabetes mellitus. Cochrane Database of Systematic Reviews 2017, Issue 12. Art. No.: CD003054. DOI: 10.1002/14651858.CD003054.pub4. Accessed 06 May 2024.

[4] Merlotti, C.; Morabito, A.; Ceriani, V.; Pontiroli, A.E. Prevention of type 2 diabetes in obese at-risk subjects: Asystematic review and meta-analysis. Acta Diabetol. 2014, 51, 853–863.

[5] Schellenberg, E.S.; Dryden, D.M.; Vandermeer, B.; Ha, C.; Korownyk, C. Lifestyle interventions for patients with and at risk for type 2 diabetes: A systematic review and meta-analysis. Ann. Intern. Med. 2013, 159, 543–551.

[6] Lende, M., & Rijhsinghani, A. (2020). Gestational Diabetes: Overview with Emphasis on Medical Management. International journal of environmental research and public health, 17(24), 9573. https://doi.org/10.3390/ijerph17249573

[7] Seuring, T., Archangelidi, O., & Suhrcke, M. (2015). The Economic Costs of Type 2 Diabetes: A Global Systematic Review. PharmacoEconomics, 33(8), 811–831. https://doi.org/10.1007/s40273-015-0268-9

[8] GBD 2019 Blindness and Vision Impairment Collaborators, & Vision Loss Expert Group of the Global Burden of Disease Study (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. The Lancet. Global health, 9(2), e144–e160. https://doi.org/10.1016/S2214-109X(20)30489-7

[9] Oh, K., Kang, H. M., Leem, D., Lee, H., Seo, K. Y., & Yoon, S. (2021). Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. Scientific reports, 11(1), 1897. https://doi.org/10.1038/s41598-021-81539-3

[10] Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35:556–64.

[11] Li, H., Liu, X., Zhong, H., Fang, J., Li, X., Shi, R., & Yu, Q. (2023). Research progress on the pathogenesis of diabetic retinopathy. BMC ophthalmology, 23(1), 372. https://doi.org/10.1186/s12886-023-03118-6

[12] Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

[13] Uddin, M. A., Islam, M. M., Talukder, M. A., Hossain, M. A. A., Akhter, A., Aryal, S., & Muntaha, M. (2024). Machine learning based diabetes detection model for false negative reduction. Biomedical Materials & Devices, 2(1), 427-443.

[14] International Diabetes Federation. (n.d.). Diabetes facts & figures. Retrieved July 20, 2024, from https://idf.org/about-diabetes/diabetes-facts-figures/

[15] Wu, J. H., Liu, T. A., Hsu, W. T., Ho, J. H. C., & Lee, C. C. (2021). Performance and limitation of machine learning algorithms for diabetic retinopathy screening: meta-analysis. Journal of medical Internet research, 23(7), e23863.

[16] Thideai. (2023). AI in Healthcare: Predictive Analytics for Disease Prevention. https://thideai.com/ai-in-healthcare-predictive-analytics-for-disease-prevention

[17] Ishaq, M., et al. (2023). Predictive model and feature importance for early detection of type II diabetes mellitus. Translational Medicine Communications. https://transmedcomms.biomedcentral.com/articles/10.1186/s41231-023-00183-4)

[18] Darmadi, D., Gardanova, Z. R., Mikhailova, M. V., Al-Qaim, Z. H., Kostyrin, E. V., Kosov, M. E., & Vasiljeva, M. V. (2023). Enhancing Global Health System Resilience and Sustainability Post-COVID-19: A Grounded Theory Approach. *Emerging Science Journal*, *7*(6), 2022-2049.

[19] Lampezhev, A. H., Kuklin, V. Z., Chervyakov, L. M., & Tatarkanov, A. A. (2023). Development and Algorithmization of a Method for Analyzing the Degree of Uniqueness of Personal Medical Data. *HighTech and Innovation Journal*, *4*(1), 122-133.

[20] Muthaiyah, S., Singh, V. A., Zaw, T. O. K., Anbananthen, K. S., Park, B., & Kim, M. J. (2023). A Binary Survivability Prediction Classification Model towards Understanding of Osteosarcoma Prognosis. *Emerging Science Journal*, *7*(4), 1294-1314.

[21] Duong-Trung, N., Hoang, X. N., Tu, T. B. T., Minh, K. N., Tran, V. U., & Luu, T. D. (2019, November). Blueprinting the workflow of medical diagnosis through the lens of machine learning perspective. In 2019 International Conference on Advanced Computing and Applications (ACOMP) (pp. 23-26). IEEE.

[22] Mustafa, T. (n.d.). Diabetes prediction dataset. Kaggle. Retrieved May 22, 2024, from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[23] Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, *122*, 56-69.

[24] Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20 - 28. https://doi.org/10.38094/jastt20165

[25] Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360-3379.

[26] Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. *Information fusion*, *6*(1), 63-81.

[27] Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science*, *1*(2), 111-117.

[28] Obaid, H. S., Dheyab, S. A., & Sabry, S. S. (2019, March). The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In *2019 9th annual information technology, electromechanical engineering and microelectronics conference (iemecon)* (pp. 279-283). IEEE.

[29] Iliou, T., Konstantopoulou, G., Ntekouli, M., Lymberopoulos, D., Assimakopoulos, K., Galiatsatos, D., & Anastassopoulos, G. (2016). Machine learning preprocessing method for suicide prediction. In *Artificial Intelligence Applications and Innovations: 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016, Thessaloniki, Greece, September 16-18, 2016, Proceedings 12* (pp. 53-60). Springer International Publishing.

[30] Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.

[31] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429-449.

[32] Tougui, I., Jilbab, A., & El Mhamdi, J. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare informatics research*, *27*(3), 189.

[33] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316.

[34] Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. *Procedia Computer Science*, *89*, 117-123.

[35] Seliya, N., Khoshgoftaar, T. M., & Van Hulse, J. (2009, November). A study on the relationships of classifier performance metrics. In *2009 21st IEEE international conference on tools with artificial intelligence* (pp. 59-66). IEEE.

[36] Alam, M. S., Ferdous, M., & Neera, N. S. (2024). Enhancing Diabetes Prediction: An Improved Boosting Algorithm for Diabetes Prediction. International Journal of Advanced Computer Science & Applications, 15(5).

[37] Gaur, A., Ray, A. K., & Saxena, S. J. (2024). Diabetes Prediction using Supervised Machine Learning. *Smart Engineering Technology and Management*, 111.

[38] Alshenawy, F. Y., & Almetwally, E. M. (2023). A COMPARATIVE STUDY OF STATISTICAL AND INTELLIGENT CLASSIFICATION MODELS FOR PREDICTING DIABETES. *Advances and Applications in Statistics*, 88(2), 201-223.

APPENDICES

Appendix A – Section III, C, 2:

GitHub Coding Repository

https://github.com/samer-glitch/Predicting-Diabetes-and-Assessing-Risk-levels-for-retinopathy-disease-Using-ML

Appendix B – Section III, E, 2 and IV

NEW CREATED DATASET

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level
Female	80	0	1	never	26	6.6	142
Female	54	0	0	No Info	28	6.6	80
Male	28	0	0	never	27	5.7	159
Female	36	0	0	current	23.45	5	155
Male	76	1	1	current	20	4.8	155
Female	20	0	0	never	27.32	6.6	85
Female	90	0	0	ever	35	9	250
Female	24	1	1	No Info	27.32	6.6	80
Male	37	0	0	never	44	7	230
Female	44	0	1	current	23.45	5	300
Male	30	1	0	never	23	6	100
Female	26	1	1	ever	26	6	199
Female	26	1	1	current	30	6	160
Female	26	1	1	No Info	26	6	250
Female	26	0	1	ever	26	6	160
Female	27	1	0	former	33	6	130
Female	33	0	0	never	23	5	120
Female	36	0	0	never	16	6	80
Female	39	0	1	No Info	30	7.5	210
Female	56	1	1	ever	27	8	200
Male	30	0	0	never	20	6	80
Male	17	1	1	never	18	7	90
Male	6	1	1	current	23	5	155
Male	14	1	0	current	26	6	180

Appendix C – Section IV, 2, a:

This appendix provides supplementary data and experimental results aimed at identifying optimal techniques, values, and parameters for the diabetes prediction program. The details outlined below encompass a variety of approaches and the corresponding performance metrics.

Experiment 1: Application of SMOTE for Class Imbalance

- Methodology: Synthetic Minority Over-sampling Technique (SMOTE) was utilized to address class imbalance.
- Results: Performance metrics for various classifiers are presented in the table below:

TABLE A1

	Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)	Confusion Matrix
0	Logistic Regression	0.879875	0.413080	0.860259	0.558148	0.954829	4.935908	[[15461, 2073], [237, 1459]]
1	Random Forest	0.946594	0.667669	0.785377	0.721756	0.968865	882.968163	[[16871, 663], [364, 1332]]
2	XGBoost	0.964379	0.841785	0.734080	0.784252	0.975445	45.254451	[[17300, 234], [451, 1245]]
3	Voting Classifier	0.955902	0.739008	0.772995	0.755620	0.969480	15.012164	[[17071, 463], [385, 1311]]

Experiment 2: Application of ADASYN (Adaptive Synthetic Sampling)

- Methodology: ADASYN was applied to generate synthetic samples adjacent to hard-to-classify minority class samples.
- Results: The performance metrics are as follows:

TABLE A2

	Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)
0	Random Forest	0.964025	0.863055	0.703250	0.775000	0.966764	330.501986
1	XGBoost	0.966186	0.895981	0.697118	0.784138	0.971217	5.990814
2	Neural Network	0.945930	0.673841	0.748620	0.709265	0.962576	3595.085286
3	Logistic Regression	0.835629	0.340054	0.920294	0.496609	0.957930	2.282274

Experiment 3: Application of BorderlineSMOTE

- Methodology: BorderlineSMOTE was used to focus on generating synthetic samples near the borderlines of class distributions.
- Results: The results are shown below:

TABLE A3

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)	Confusion Matrix	
0	Logistic Regression	0.860062	0.375656	0.886203	0.527646	0.953037	2.844599	[[15036, 2498], [193, 1503]]
1	Random Forest	0.936453	0.605804	0.800118	0.689533	0.968058	950.399149	[[16651, 883], [339, 1357]]
2	XGBoost	0.952522	0.714286	0.769458	0.740846	0.975019	43.765607	[[17012, 522], [391, 1305]]
3	Voting Classifier	0.946854	0.664390	0.803066	0.727176	0.968929	17.103075	[[16846, 688], [334, 1362]]

Staying on SMOTE emerged as the best option since it outperforms all other techniques.

Appendix D – Section IV, 2, b

A second approach was explored by changing the prediction threshold for ML models to achieve higher Recall values, which is a crucial metric in the context of diabetes predictions.

The results for normal predictions are as follows:

TABLE B0

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)	Confusion Matrix	
0	Logistic Regression	0.879875	0.413080	0.860259	0.558148	0.954829	4.935908	[[15461, 2073], [237, 1459]]
1	Random Forest	0.946594	0.667669	0.785377	0.721756	0.968865	882.968163	[[16871, 663], [364, 1332]]
2	XGBoost	0.964379	0.841785	0.734080	0.784252	0.975445	45.254451	[[17300, 234], [451, 1245]]
3	Voting Classifier	0.955902	0.739008	0.772995	0.755620	0.969480	15.012164	[[17071, 463], [385, 1311]]

Experiment 4: Adjusting the Prediction Threshold to 0.6

- Methodology: The prediction threshold was raised to 0.6 to evaluate the impact on precision and recall, prioritizing the reduction of false positives which is crucial for clinical diagnostics where confirming the presence of diabetes is critical.

- Results: The performance metrics with this higher threshold are detailed in the appendix. This adjustment typically resulted in higher precision but slightly lower recall, indicating fewer false positives at the expense of missing some true positives.

TABLE B1

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)	Confusion Matrix	
0	Logistic Regression	0.906344	0.481714	0.815448	0.605649	0.954829	0.264111	[[16046, 1488], [313, 1383]]
1	Random Forest	0.958658	0.779640	0.740566	0.759601	0.968689	26.816174	[[17179, 355], [440, 1256]]
2	XGBoost	0.970203	0.936965	0.709906	0.807783	0.975445	1.323363	[[17453, 81], [492, 1204]]
3	Voting Classifier	0.962611	0.822868	0.734080	0.775943	0.970382	26.425162	[[17266, 268], [451, 1245]]

Experiment 5: Adjusting the Prediction Threshold to 0.25

- Methodology: The threshold was lowered to 0.25 to maximize recall. This approach is aimed at reducing false negatives, essential for early screening where capturing as many potential cases as possible is more critical than the precision of each prediction.
- Results: This lower threshold significantly improved recall but at the cost of precision, as detailed in the appendix. The increase in recall makes this threshold suitable for preliminary screenings.

TABLE B2

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)	Confusion Matrix	
0	Logistic Regression	0.781539	0.280225	0.941627	0.431913	0.954829	0.329488	[[13432, 4102], [99, 1597]]
1	Random Forest	0.886271	0.430077	0.890330	0.579988	0.968202	26.290070	[[15533, 2001], [186, 1510]]
2	XGBoost	0.917681	0.519584	0.883844	0.654442	0.975445	2.742484	[[16148, 1386], [197, 1499]]
3	Voting Classifier	0.857046	0.375385	0.935142	0.535720	0.970287	25.459364	[[14895, 2639], [110, 1586]]

Experiment 6: Adjusting the Prediction Threshold to 0.4

- Methodology: A moderate threshold adjustment to 0.4 was tested to find a balance between recall and precision. This setting aims to maintain a reasonable rate of true identifications while controlling the number of false positives.
- Results: The results, as detailed in the appendix, show that this threshold offers a balanced trade-off, making it a potentially viable option for contexts where both identifying cases and maintaining precision are important.

TABLE B3

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Running Time (s)	Confusion Matrix	
0	Logistic Regression	0.849922	0.359338	0.896226	0.512994	0.954829	0.263994	[[14824, 2710], [176, 1520]]
1	Random Forest	0.927665	0.560878	0.828420	0.668888	0.968033	25.806485	[[16434, 1100], [291, 1405]]
2	XGBoost	0.954082	0.718431	0.788325	0.751757	0.975445	1.291959	[[17010, 524], [359, 1337]]
3	Voting Classifier	0.920749	0.531387	0.858491	0.656447	0.970437	25.257277	[[16250, 1284], [240, 1456]]