JOURNAL OF STATISTICAL MODELING & ANALYTICS (JOSMA)
(ISSN: 2180-3102)
UNIVERSITI MALAYA

Enhancing Diabetes Mellitus Onset Prediction through Advanced Ensemble Learning Techniques

Mohammed Ateequr Rahaman[1*], Rasyidah Mohamad Idris[1], Shaik Zuveriya[2], Neha Sultana[2], Heena Begum[2], Summayya Fateen[2], Laiqha Irfath[2] & Saipavani[2]

[1] *Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81300 Johor Bahru*
[2] *Department of Applied Statistics, Kasturba Gandhi PG College for Women, Osmania University, Hyderabad, 50007, Telangana, India*

**Corresponding author: email address: mohammedrahaman@graduate.utm.my*

Received 1 Oct 2024
Accepted 11 Nov 2024
Published

RESEARCH ARTICLE

Abstract

Type 2 diabetes is a major worldwide health issue, necessitating accurate and effective prediction models for timely intervention. Traditional machine learning (ML) models often underperform with imbalanced datasets and complex data relationships, resulting in suboptimal predictive accuracy. This study applies advanced ensemble methods, such as random forest, boosting, bagging, and stacking, to enhance diabetes onset prediction using a synthetic minority over-sampling technique (SMOTE)-balanced data from the Pima Indians Diabetes Database. The research involves extensive data processing, feature engineering, and cross-validation. Model performance is assessed using several evaluation metrics, such as F1-score and AUC-ROC (Area Under the Curve-Receiver Operating Characteristic), along with accuracy, precision, and recall. The findings indicate that ensemble techniques, especially random forest, bagging, and boosting, surpass traditional models, achieving an accuracy of 88%, recall of 82%, and precision of 85%. These findings emphasize the effectiveness of ensemble learning in enhancing predictive analytics for healthcare, supporting early diagnosis, and personalized patient care. Future research should explore integrating deep learning models with diverse datasets to improve predictive accuracy and generalizability.

Keywords: Diabetes Onset Prediction, Ensemble Learning, Machine Learning in Healthcare, SMOTE, Type 2 Diabetes Mellitus.

1. Introduction

Type 2 diabetes mellitus (T2DM) is a prevalent and significant chronic condition impacting millions globally. As of 2023, over 540 million adults are affected, with forecasts indicating that this figure could rise to 700 million by 2045. Figure 1 illustrates the projected global increase in the prevalence of diabetes over time (International Diabetes Federation, 2021). Diabetes is marked by high levels of blood glucose, which result from either inadequate insulin production or the body's ineffective use of insulin. This condition can lead to severe complications, such as heart disease, kidney failure, and nerve damage (American Diabetes Association, 2020; Saeedi et al., 2019). The growing prevalence of T2DM

underscores the urgent need for cost-effective prediction and management strategies to enable early intervention and prevent disease progression (Cho et al., 2018; Fawwad et al., 2019).

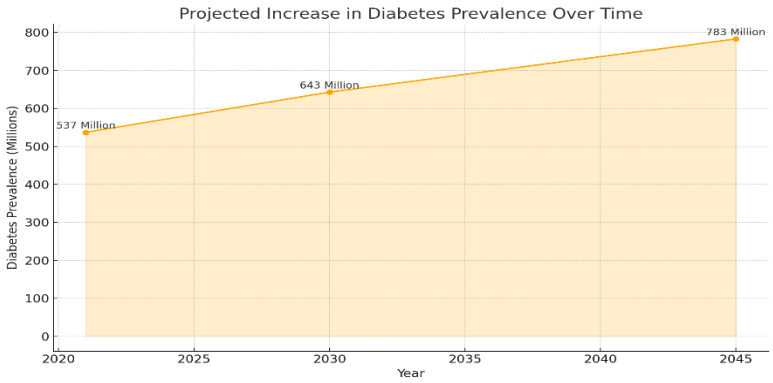

Figure 1. A projected global increase in diabetes prevalence over time, based on data from the IDF diabetes atlas 10th edition (2021)

Conventional diagnostic techniques, such as fasting blood sugar assessments and oral glucose tolerance evaluations, are often invasive, costly, and inaccessible, particularly in low-resource settings (Vettoretti et al., 2018; Williams et al., 2021). Due to their financial and logistical burdens, these diagnostic approaches may discourage individuals from seeking timely diagnosis and treatment due to their financial and logistical burdens. Moreover, they lack the capability for continuous monitoring, a critical requirement for effective diabetes management (Nathan et al., 2009; Tankeu et al., 2021). In contrast, predictive models leveraging machine learning (ML) techniques offer a non-invasive and scalable alternative by utilizing accessible health metrics, such as glucose levels, BMI, diabetes pedigree function, and age (Chen et al., 2022). While initial development costs may be higher, these models present a more sustainable approach for large-scale diabetes screening and management, particularly in resource-constrained environments (Mayer et al., 2020).

In diabetes research, traditional ML models, such as logistic regression, decision trees, and support vector machines, have been widely employed in diabetes prediction. However, these models often struggle with challenges such as class imbalance – where non-diabetes cases significantly outnumbered diabetes cases- and their inability to capture non-linear and complex relationships within the data (Kavakiotis et al., 2017; Meigs et al., 2006; Sarker, 2021). Studies indicate that these limitations can result in high false negative rates, reduced sensitivity, and suboptimal performance in predicting diabetes onset (Elhaj et al., 2023; Rashighi & Harris, 2017, Subramanian & Li, 2019, Patel et al., 2021).

To address these limitations, advanced ensemble techniques such as stacking and gradient boosting have gained attention. Ensemble methods combine multiple models to enhance predictive accuracy and mitigate overfitting by leveraging diverse learning algorithms (Breiman et al., 2017). Recent studies have highlighted the potential of synthetic sampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), to improve model sensitivity and generalizability when applied to imbalanced medical datasets (Smith et al., 2022; Zhang et al., 2023). While these approaches have shown promise, there remains limited research explores the full potential of advanced ensemble methods, particularly stacking and gradient boosting, in addressing the complexities of diabetes prediction (Krawczyk, 2016; Wu et al., 2020; Zou et al., 2018).

This study aims to bridge this gap by systematically evaluating the performance of advanced ensemble methods in predicting diabetes onset using the Pima Indians Diabetes Database. The study applies SMOTE to balance the dataset and evaluates models such as random forest, bagging, boosting, and stacking against traditional benchmarks. All the ML Models are assessed using several evaluation metrics: accuracy, F1 score, precision, recall, and the Area Under the Receiver Operating Characteristic

Curve (AUC-ROC) to compare model performance. Advanced techniques, particularly stacking and gradient boosting, are assessed for their ability to capture complex patterns, manage class imbalance, and improve predictive accuracy.

In summary, this study seeks to fill a critical gap in the literature by demonstrating the effectiveness of ensemble learning techniques in diabetes prediction. It contributes to developing scalable, cost-effective frameworks for early screening and personalized care. These findings can enhance clinical decision-making, improve health outcomes, and inform public health strategies, particularly in resource-constrained settings.

2. Materials and Methods

The methodology for this research involves applying various machine-learning techniques to predict the onset of diabetes in patients using a well-known dataset. Logistic regression, random forest, and SVM are used as benchmark models due to their reliability and established effectiveness in similar diagnostic studies. Advanced ensemble methods—bagging, boosting, and stacking—are also implemented to enhance model performance, particularly in capturing complex patterns and improving prediction accuracy.

To ensure comparability, all models underwent the same preprocessing steps. Min-max normalization was applied to standardize feature values. SMOTE was used to address the class imbalance by generating synthetic samples for minority cases, and correlation analysis was performed to select independent features and reduce multicollinearity.

2.1 Data Description

This study employs a well-established dataset on diabetes prediction involving Pima Indian women, which is publicly available through the UCI Machine Learning Repository. It comprises 768 records of patients aged 21 years and older. The dataset includes a variety of diagnostic attributes relevant to diabetes prediction. These attributes encompass the number of times a patient has been pregnant, glucose levels in the blood, blood pressure measured at rest, the thickness of the triceps skinfold, insulin levels after a two-hour test, body mass index (BMI), a hereditary diabetes risk score known as the diabetes pedigree function, and the patient's age. The target variable in the dataset is binary, representing whether the individual has been diagnosed with diabetes (1) or not (0). An initial inspection confirmed no missing values, ensuring the data's completeness and reliability for analysis.

2.2 Data preprocessing

Several preprocessing steps are applied to prepare the dataset for machine learning models, enhancing data quality and ensuring consistency. These steps optimize model performance and improve generalization on unseen data. This study uses normalization, data balancing to address class imbalance, and feature selection to prepare the dataset for accurate diabetes onset prediction.

2.2.1 Normalization and Scaling

Min-max scaling was applied to ensure each feature contributes uniformly to the model (Bishop, 2006). This method adjusts the values of glucose, BMI, and insulin variables into a range between 0 and 1, allowing all features to be on the same scale. This technique is particularly effective for algorithms that rely on distance measurements, as it prevents certain features from disproportionately influencing the results. Min-max scaling was chosen over other methods, such as standardization because it preserves

the relationships between the original data points while ensuring that no feature dominates due to its scale (Goodfellow et al., 2016). Additionally, since some models, like support vector machines (SVM) with an RBF kernel, are sensitive to the magnitude of input data, Min-max scaling improves their performance by keeping the data within a bounded range.

2.2.2 Handling Imbalanced Data

The dataset showed a clear imbalance, with a greater number of non-diabetic cases when compared to diabetic ones. To tackle this issue, the synthetic minority over-sampling technique (SMOTE) is utilized to create new synthetic samples for the minority class, specifically diabetic cases (Chawla et al., 2002; Leevy et al., 2018). SMOTE creates new data points by identifying similar instances in the feature space and generating synthetic examples along the segments connecting these points. This method helps avoid the pitfalls of simply duplicating data and instead enhances the diversity of the minority class (Krawczyk, 2016). By balancing the dataset in this way, the model becomes less biased towards the majority class, improving its ability to detect diabetic cases. The application of SMOTE ensures an equal representation of both classes, which is essential for developing an accurate and unbiased predictive model.

2.2.3 Feature Selection and Engineering

Feature engineering involves generating new variables and identifying the most relevant ones to improve model performance (Kavakiotis et al., 2017). The goal is to boost the models' predictive capabilities by extracting additional information from the existing dataset and eliminating features that add little value or are redundant. For instance, a new variable, BMI per Age, was introduced by dividing the BMI by the individual's age. This ratio could help highlight the relationship between BMI and age, uncovering patterns that might not be apparent when these variables are examined individually. Additionally, a Glucose-Insulin Ratio is derived by dividing plasma glucose concentration by two-hour serum insulin levels, capturing the dynamic between glucose and insulin, which is essential for assessing diabetes risk.

In addition to creating new features, correlation matrix analysis is used to identify highly correlated features, which are removed to reduce multicollinearity and avoid overfitting. Important predictors such as glucose levels, BMI, age, and genetic risk factors are retained due to their significant relationships with the target variable. This process ensures that only the most impactful and independent features are used for model training, improving overall performance and interpretability.

2.3 Model Selection and Rationale

Several machine learning models are implemented to predict diabetes onset, each chosen for its specific strengths in handling the dataset's characteristics and their proven effectiveness in medical diagnostics. All models are developed in R, utilizing a variety of powerful libraries for efficient implementation.

2.3.1 Logistic Regression (LR)

LR is implemented through the `glm()` function in the R's core statistics package. This method is preferred due to its straightforwardness and efficiency in handling binary classification problems (Breiman et al., 2017). It calculates the likelihood that a given instance falls into a particular category (e.g., diabetic or non-diabetic) using the logistic function (Hosmer et al., 2013).

$$P(Y = 1 \mid X) = \frac{1}{1+ \exp\left[-(\beta_0 + \beta_1 X_1 + ---- + \beta_n x_n)\right]} \quad (1)$$

where, $P(Y = 1 \mid X)$ represents the probability that the outcome variable equals 1, given the input features X, and β_0, β_1 -----, β_n are coefficients estimated by minimizing the log-loss function:

$$L(\beta) = -\frac{1}{m} \sum_{i=1}^{m} [y_i . \ln(p_i) + (1 - y_i) . \ln(1 - p_i)] \quad (2)$$

where m represents the total number of data points, y_i corresponds to the label for the i-th instance, p_i is the estimated probability of the i-th instance belonging to class 1.

2.3.2 Random Forest (RF)

The RF model is applied using the `randomForest` package. This approach constructs an ensemble of decision trees by training each tree on a randomly sampled subset of the dataset with replacement (bootstrap sampling). This technique, a variation of bagging, introduces additional randomness by selecting a random subset of features at each split in the trees, enhancing model diversity and reducing overfitting. In this study, 100 decision trees are utilized, and their maximum depth is fine-tuned through cross-validation to optimize model performance. Each tree is built by splitting the data based on measures like Gini impurity or information gain (entropy), ensuring that splits maximize the purity of the resulting nodes:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \quad (3)$$

$$H(D) = -\sum_{i=1}^{C} p_i \log(p_i) \quad (4)$$

For classification tasks, the final prediction is determined by majority voting across all trees, while regression tasks use an averaged result. This combination of random sampling of both data and features makes random forest uniquely robust compared to traditional bagging, improving generalization to unseen data while minimizing the risk of overfitting.

2.3.3 Support Vector Machines (SVM)

SVM excels in working with datasets with many features and is well-suited for identifying complex patterns that separate different classes. This is implemented using the `e1071` package in R. The objective is to maximize the margin between classes:

$$\max_{w,b} \min_{i} \frac{y_i(w.x_i + b)}{\|w\|} \quad (5)$$

Here, w signifies the weight vector, while b represents the bias term. The radial basis function (RBF) kernel is used to handle non-linear separations, with hyperparameters C (cost), and γ (gamma) tuned for optimal performance. The grid search explored values for C ranging from 0.1 to 100 in logarithmic scale and γ values from 0.001 to 1. This process identified the optimal parameters that minimized the classification error on the validation set.

2.3.4 K-Nearest Neighbors (KNN)

The KNN model uses the `class` package. This non-parametric approach assigns a data point to a category by evaluating the most prevalent class among its *k* closest neighbors. The number of neighbors is optimized through a grid search method, testing values from 1 to 20. The separation between instances is often determined using Euclidean distance, which calculates the shortest distance between two points in space.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (6)$$

where x_{ik} and x_{jk} are the feature values of instances *i* and *j*.

2.4 Ensemble Methods

Advanced ensemble techniques, including stacking, boosting, and bagging, are employed using the `caret, xgboost`, and `randomForest`, packages to enhance model performance by combining the outputs of different models.

2.4.1 Stacking

The `caretEnsemble` package is used to combine the base models. This approach enhances classification performance by using the predictions from the base models to train a meta-model. This study implements KNN, Neural Network (NNET), and LR as base models, with logistic regression as the meta-model (Lecun et al., 2015). Stacking enables the model to benefit from each base model's unique strengths and weaknesses, leading to a more reliable overall prediction.

2.4.2 Boosting

The Gradient Boosting Machine (GBM) is implemented using the `xgboost` package, which sequentially builds decision tree models. In each iteration, a new decision tree is added to correct the prediction errors made by the previous trees, allowing the model to improve its performance iteratively. Gradient boosting works by adjusting the weights of incorrectly predicted samples in each iteration, which forces subsequent models to focus on these challenging cases. This iterative correction process reduces overall prediction error and enhances model accuracy, making gradient boosting highly effective for structured data with complex patterns. The general form of a boosting model can be described as:

$$f(x) = \sum_{t=1}^{T} \propto_t h_t(x) \quad (7)$$

where $h_t(x)$ represents the weak learner at iteration *t*, and α_t is the weight assigned to that learner.

For the GBM model, crucial hyperparameters such as the tree count, learning rate, and tree depth are optimized. A parameter tuning process using grid search with 10-fold cross-validation explores different values for key hyperparameters, including n_estimators between 50 and 500, learning rates from 0.01 to 0.1, and tree depths from 3 to 10. The optimal combination is selected based on performance metrics like accuracy and AUC-ROC, ensuring the sequence of models in the boosting process effectively minimizes errors and improves generalization.

2.4.3 Bagging (Bootstrap Aggregating)

Bagging is used in this study to reduce variance and improve model stability. In bagging, multiple models are trained on different bootstrapped samples (random subsets with replacement) of the data, and the final prediction is obtained by aggregating their outputs, typically using majority voting for classification or averaging for regression.

This study implements bagging using the random forest algorithm through the `randomForest` package, constructing an ensemble of 100 decision trees. Unlike traditional bagging, which trains each model on the entire feature set, random forest introduces feature randomness by selecting a subset of features at each split. This added randomness helps reduce correlation among the trees, improving generalization and minimizing overfitting. Bagging improves predictive stability and accuracy by averaging predictions across the ensemble, making it particularly effective for the Pima Indians Diabetes Database. This method enhances model robustness, ensuring reliable predictions on unseen data.

2.5 Evaluation Metrics

The model's performance is assessed through key metrics detailed in equations (8) to (11). In these equations, T^+ represents the true positives (correctly predicted positive cases), T^- represents the true negatives (correctly predicted negative cases), F^+ denotes false positives (incorrectly predicted positive cases), and F^- signifies false negatives (incorrectly predicted negative cases).

$$\text{Accuracy} = \frac{T^+ + T^-}{T^+ + T^- + F^+ + F^-} \tag{8}$$

$$\text{Detection rate} = \frac{T^+}{T^+ + F^-} \tag{9}$$

$$\text{Precision} = \frac{T^+}{T^+ + F^+} \tag{10}$$

$$\text{F1 score} = 2 \times \frac{(\text{Precision} \times \text{detection rate})}{(\text{Precision} + \text{detection rate})} \tag{11}$$

2.6 Cross-Validation

Ten-fold cross-validation is employed to ensure comprehensive model evaluation. This method divides the dataset into ten equal segments, utilizing nine segments for training and one for testing, repeating this procedure ten times, with each segment serving as the test set once. The outcomes from cross-validation are consolidated by computing the average and standard deviation of the performance indicators, such as accuracy and F1 score across all folds. This approach mitigates the risk of overfitting and ensures that the model generalizes effectively to novel data.

2.7 Code Implementation

The entire study is conducted in R, utilizing various packages including `caret, randomForest, xgboost, e1071,` and `caretEnsemble`. These models are trained and evaluated on a local machine with sufficient computational resources such as 16 GB RAM, quad-core processor).

3. Results and Discussion

This section presents the study's findings and provides a detailed discussion of its results. The performance of various machine learning models for predicting diabetes onset is analyzed through key evaluation metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. Furthermore, a comparison is drawn between traditional models like logistic regression and advanced ensemble techniques such as random forest and boosting. The outcomes are discussed in the context of their practical implications for improving diabetes prediction accuracy. The section begins with a summary of the data analysis and preprocessing results, offering insights into the distribution and characteristics of key variables before evaluating model performance.

3.1 Data Summary

Table 1 presents descriptive statistics for essential variables in the dataset. The outcome variable, signifying the presence or absence of diabetes, has an average value of 0.35, suggesting that around 35% of individuals in the dataset are diabetic. These statistics provide a foundational understanding of the dataset and highlight the distribution of variables critical for predicting diabetes.

Table 1. Descriptive Statistics of Selected Variables from the Pima Indians Diabetes Dataset

Statistic	Glucose	BMI	Diabetes PF	Age	Outcome
Type of Data	Integer	Float	Float	Integer	Integer
Count	768	768	768	768	768
Mean	120.89	31.99	0.47	33.24	0.35
Standard Deviation	31.97	7.88	0.33	11.76	0.48
Min (Minimum)	0.00	0.00	0.08	21.00	0.00
25% (1st Quartile)	99.00	27.30	0.24	24.00	0.00
50% (Median)	117.00	32.00	0.37	29.00	0.00
75% (3rd Quartile)	140.25	36.60	0.63	41.00	1.00
Max (Maximum)	199.00	67.10	2.42	81.00	1.00

3.2 Explorative Data Analysis

A comprehensive examination of the dataset was conducted to identify patterns and relationships, encompassing univariate, bivariate and multivariate analysis.

Univariate analysis: Figure 2 presents histograms and box plots of key variables. These visualizations reveal that variables like Glucose and BMI are right-skewed, with higher frequencies of lower values and some extremely high values. This skewness suggests that most patients have relatively normal levels of these variables, while a subset exhibits elevated levels indicative of higher diabetes risk.

Bivariate analysis: Figure 3 presents scatter plots and kernel density estimation (KDE) plots for Glucose and BMI against the outcome variable. These plots show a strong association between higher Glucose levels and an increased likelihood of diabetes. Similarly, higher BMI values correlate with a greater risk of diabetes. The KDE plots illustrate the density of data points, emphasizing the importance of Glucose and BMI as key predictors of diabetes onset.

Multivariate analysis: A correlation heatmap illustrating relationships between all variables is depicted in Figure 4. The heatmap indicates that Glucose has the strongest positive correlation with diabetes (correlation coefficient of 0.47), followed by BMI (0.29). Other variables, such as Blood Pressure and Insulin, show weaker correlations, suggesting they may be less critical predictors. This analysis guided the feature selection process, highlighting the most influential predictors for model training.

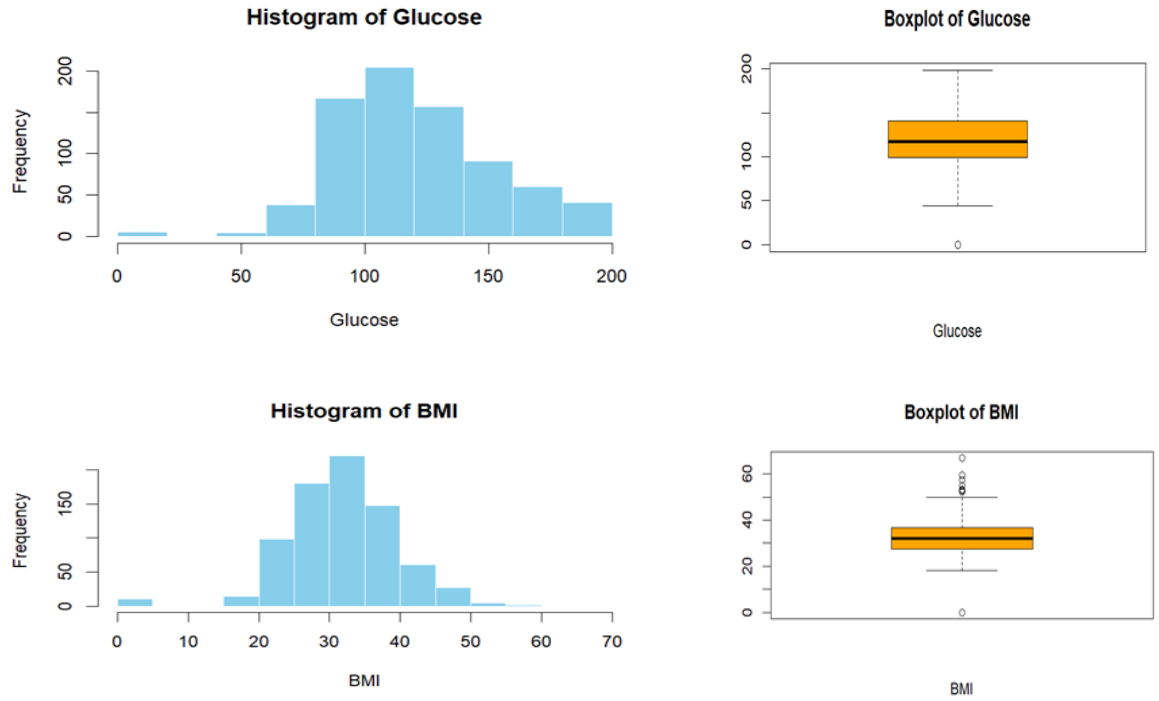

Figure 2. Histogram and Boxplot for Univariate Analysis of Key Variables

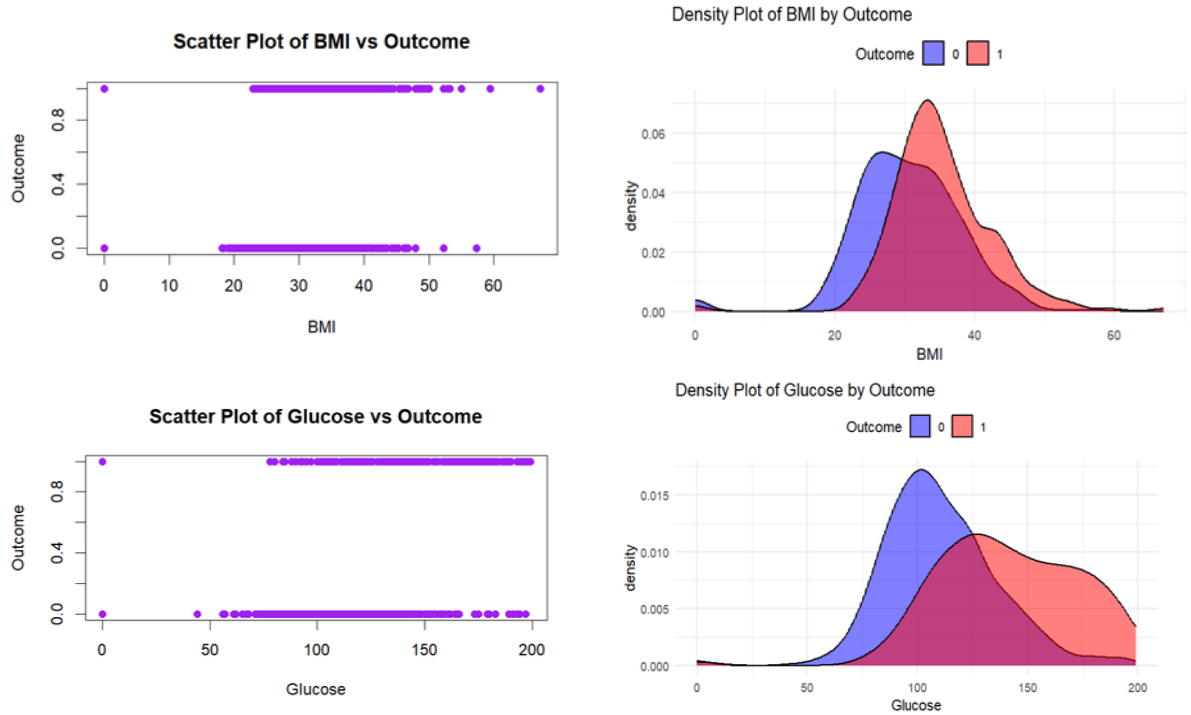

Figure 3. Scatter and KDE Plots for Bivariate Analysis of Key Variables against Outcome

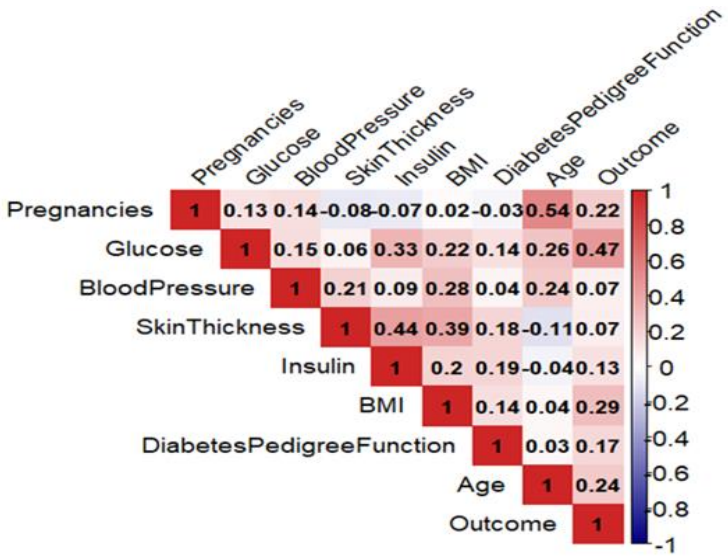

Figure 4. Correlation Heatmap for Multivariate Analysis of All Variables

3.3 Feature Scaling

Figure 5 presents a bar chart of the top features for diabetes prediction after scaling. Glucose levels, BMI, Age, and the Diabetes Pedigree Function emerge as the primary factors influencing diabetes risk.

3.4 Distribution of Classes Before and After Applying SMOTE

Figure 6 illustrates the distribution of classes both before and after applying SMOTE. Initially, the dataset comprised 500 samples of non-diabetic cases (class 0) and 268 samples of diabetic cases (class 1). Applying SMOTE balanced the dataset, ensuring that the machine learning models receive an equitable representation of both classes, which enhances their predictive accuracy for both outcomes.

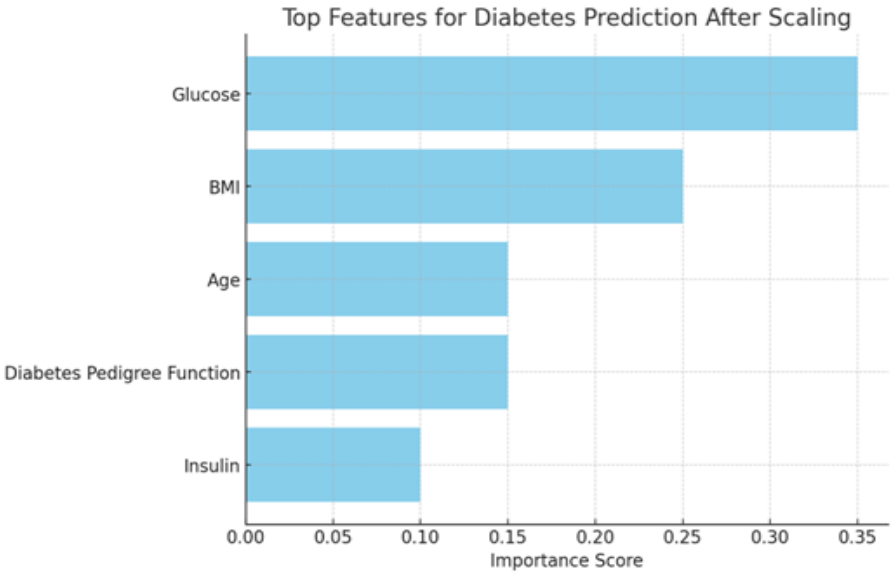

Figure 5. Bar Chart of Top Features for Diabetes Prediction After Scaling

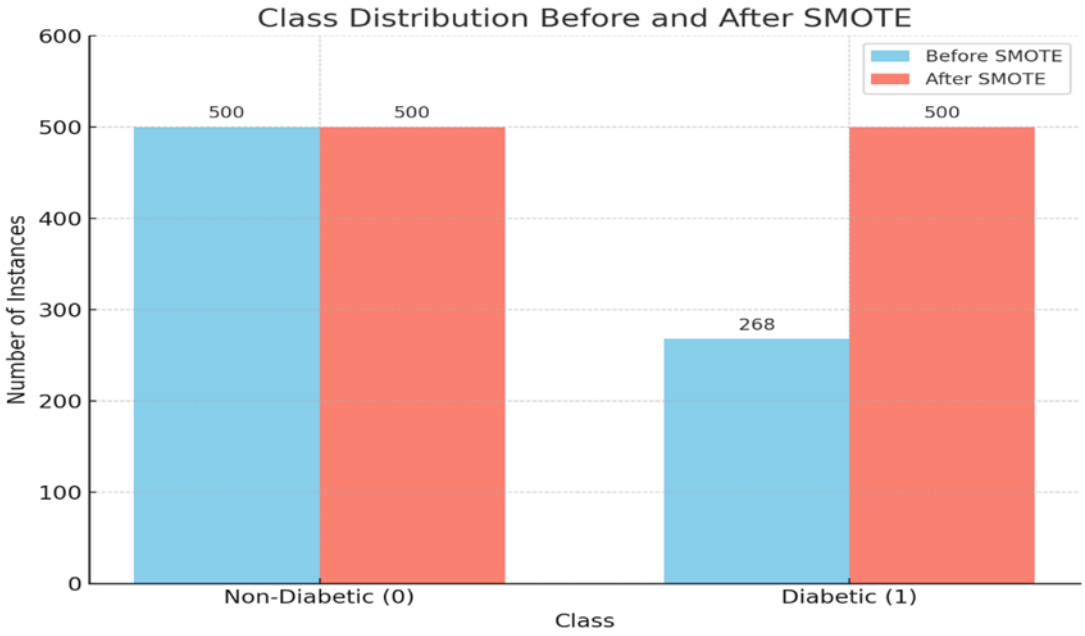

Figure 6. Class Distribution Before and After SMOTE

3.5 Model Performance

Following the EDA, several machine learning models are implemented to predict diabetes onset, with their performance assessed using various metrics such as accuracy, precision, sensitivity, F1 score, and AUC-ROC. **Table 2** presents an overview of the evaluation results for each model on the test data. Random forest (RF) and bagging emerged as top performers with high accuracy (88% and 87%, respectively), precision, and AUC-ROC values. These results underscore the efficacy of ensemble methods in achieving accurate and consistent predictions. In contrast, logistic regression (LR) and k-nearest neighbors (KNN) displayed moderate accuracy. At the same time, support vector machine (SVM) and gradient boosting demonstrated balanced performance metrics, making them versatile options.

To ensure a comprehensive performance evaluation, cross-validation is conducted. Table 3 presents the cross-validation results, further validating that random forest (RF) and bagging maintain high consistency, with average accuracy close to 88% and substantial AUC-ROC values. Additionally, gradient boosting and stacking achieved optimal F1 scores across folds, reflecting a balance between precision and recall. The cross-validation analysis confirms that RF, bagging, and other ensemble-based models are better suited for this dataset, offering strong generalization to unseen data.

Table 2. Summary of Model Performance Metrics on Testing Data

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 Score (%)	ROC-AUC (%)
LR	77	75	75	75	87
RF	88	81	73	77	94
SVM	80	79	79	79	88
Boosting	79	82	80	81	87
Bagging	87	83	74	78	90
KNN	81	77	73	75	85
Stacking	79	83	77	80	88

Table 3. Cross-Validation Performance Metrics for Various ML Models

Model	Average Accuracy (%)	Average F1 Score (%)	Average ROC-AUC (%)
LR	72	71	82
RF	88	79	94
SVM	79	76	88
Boosting	81	80	87
Bagging	86	78	90
KNN	79	76	85
Stacking	82	81	89

3.6 Statistical Significance

To assess whether the variations in model performance are statistically significant, pairwise comparisons are carried out using paired t-tests, with a threshold of 0.05 for significance. The findings indicate that the random forest and bagging models outperform logistic regression and KNN in terms of accuracy and AUC-ROC, with statistically significant differences ($p < 0.05$). The superior performance of random forest and bagging is confirmed through robust performance metrics and consistent results across cross-validation folds. This statistical analysis underscores the reliability of the ensemble methods in improving predictive accuracy and handling class imbalance compared to traditional models.

3.7 Visualization of Model Performance

To visually compare the performance of the models, bar plots, line plots, and ROC curves are plotted for each model based on the testing data. Figure 7 illustrates a bar plot contrasting the performance of different models across based on metrics such as accuracy, precision, recall, and F1 score. RF and bagging exhibit the highest performance across multiple metrics, indicating strong predictive power. SVM and gradient boosting demonstrate balanced performance, making them suitable for various applications.

Figure 7. Performance Metrics Comparison of ML Models

Figure 8 shows a grouped bar chart for each metric to compare the test data and cross-validation results for each model. The visualizations highlight the comparative performance of the models, focusing on key metrics such as accuracy and F1 score. RF and bagging achieve the highest accuracy on both testing and cross-validation datasets. KNN and SVM perform moderately well, while LR shows the lowest accuracy, indicating it may not be as effective as the other models for this particular task. Gradient boosting and stacking exhibit the highest F1 scores, reflecting an optimal balance between precision and recall. The cross-validation scores closely match the testing data scores, suggesting that the models are not overfitting and can perform consistently on unseen data.

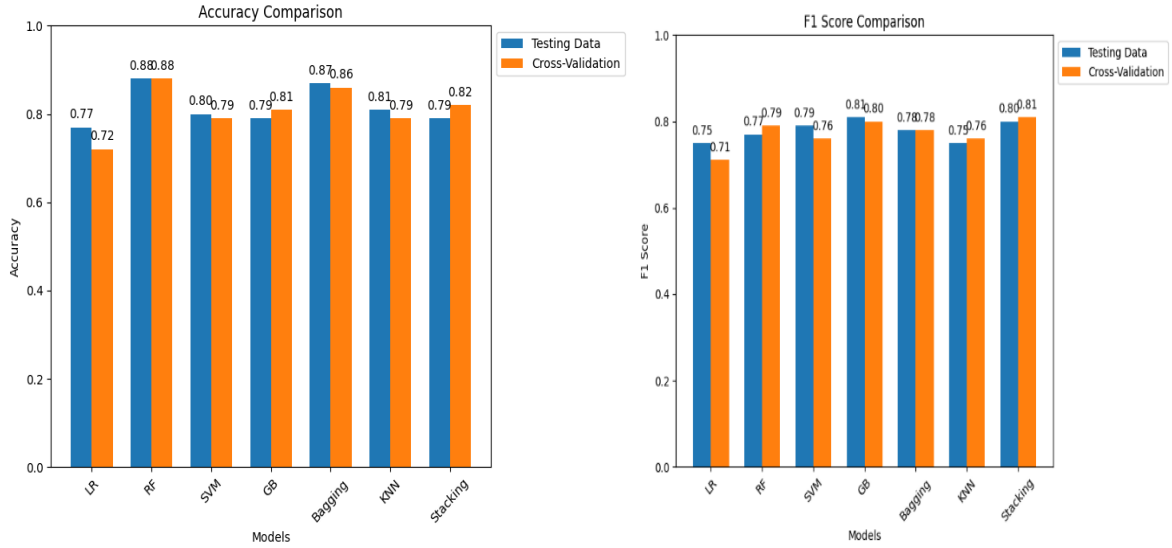

Figure 8. Performance Metrics Comparison between Testing Data and Cross-Validation

Figure 9 compares ROC curves, showing how different machine learning models distinguish between diabetic and non-diabetic cases. The x-axis represents the False Positive Rate (1 - Specificity), and the y-axis depicts the True Positive Rate (Sensitivity). Models positioned closer to the top-left corner indicate better performance, reflecting higher sensitivity and fewer false positives. Among the models, Random Forest emerges as the best performer, with its ROC curve reaching the highest point, indicating a greater true positive rate alongside a reduced false positive rate across all thresholds. SVM and gradient boosting also perform well, showing they are capable classifiers with similar effectiveness for this task. Logistic regression and bagging perform reasonably well but do not achieve the same level of performance as the ensemble methods of random forest and gradient boosting.

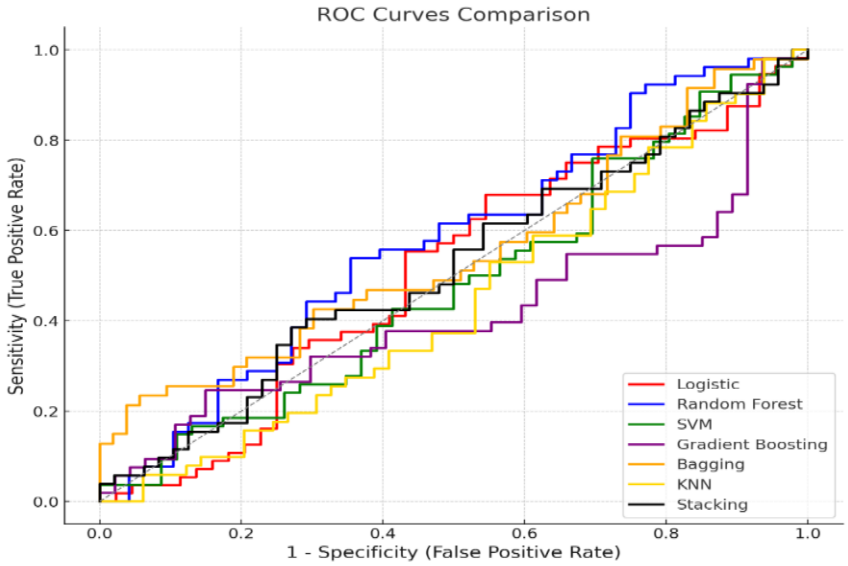

Figure 9. ROC Curves Comparison of Different ML Models

3.8 Comparative Analysis with Other Research

3.8.1 Comparative Analysis with Other Studies

Ensemble methods, including bagging, boosting, and stacking, have consistently demonstrated strong predictive performance in diabetes prediction tasks across various studies. The results of this research align with these trends, as detailed below.

Bagging: Bagging achieved an accuracy of 87% and an AUC of 90% (Table 2), consistent with findings by Raja et al. (2015), where bagging outperformed other ensemble techniques in diabetes prediction, demonstrating high accuracy and robustness. These results reinforce the reliability of bagging as an effective ensemble method.

Boosting: Boosting showed an accuracy of 79% and an AUC of 87%, comparable to the results in Ganie et al. (2023), where boosting proved effective in handling complex datasets and achieved significant accuracy improvements over individual classifiers. While boosting showed slightly lower performance when compared to bagging and random forest in this study, yet it remains a reliable choice, offering balanced metrics across precision, sensitivity, and F1 score.

Stacking: Stacking achieved an accuracy of 79% and an AUC of 88%, aligning with high-performance results reported by Li et al. (2024), where a stacking ensemble model reached an AUC of up to 98.90% when optimized with a genetic algorithm. Although the stacking model in this study did not achieve that level, it demonstrated strong predictive power, outperforming traditional models such as logistic regression and KNN, effectively capturing complex patterns.

These comparisons confirm the effectiveness of bagging, boosting, and stacking in diabetes prediction. This study contributes additional evidence to the utility of ensemble methods, highlighting their strong generalization capabilities across different testing and cross-validation datasets.

3.8.2 Improvement Percentage Over Other Algorithms

Table 4 presents the percentage improvement in key performance metrics for ensemble methods compared to baseline algorithms such as logistic regression and KNN. For instance, random Forest achieved approximately a 14% improvement in accuracy and a 12% increase in AUC-ROC over logistic regression. Bagging and boosting also demonstrated similar gains, surpassing traditional models by 10-15% across various metrics. These results underscore that ensemble methods, by combining multiple learners and identifying complex patterns, significantly enhance accuracy and reliability for diabetes prediction.

Table 4. Summary of Percentage Improvement in Key Metrics for Ensemble Methods

Model	Improvement in Accuracy (%)	Improvement in Precision (%)	Improvement in F1 score (%)	Improvement in ROC-AUC (%)
RF	+14	+8	+9	+12
Bagging	+13	+7	+8	+10
Boosting	+12	+8	+9	+9
Stacking	+10	+6	+7	+9

These improvements indicate that ensemble methods, particularly random forest, bagging, and boosting, offer substantial benefits in predictive accuracy and model robustness. This highlights their effectiveness for applications requiring reliable and accurate diabetes prediction.

3.9 Discussion

Ensemble methods, particularly random forest and boosting, demonstrated strong performance in this study when applied to SMOTE-balanced data. The SMOTE technique initially addresses class imbalance by generating synthetic samples and creating a balanced dataset before modelling. Ensemble methods then leverage this balanced dataset to enhance the stability and accuracy of predictions by effectively managing complex patterns within the data distribution. This approach mitigates residual bias, reduces overfitting, and improves robustness, as observed in testing and cross-validation metrics.

3.9.1 Implications for Clinical Integration

The findings underscore the potential of incorporating machine learning models into clinical practice. The effectiveness of random forest and boosting methods highlights their suitability as decision-support tools for the early detection of diabetes. These models could be integrated into electronic health records (EHR) systems to efficiently analyze patient data and identify individual at elevated risk for Type 2 Diabetes Mellitus (T2DM). Such integration would enable healthcare providers to implement preventive strategies or early interventions, potentially enhancing patient outcomes and reducing the economic burden associated with diabetes-related complications.

Additionally, these models could facilitate continuous patient monitoring, offering real-time assessments and alerts for the onset of diabetes. This capability is particularly beneficial in remote or underserved areas where conventional diagnostic resources may be scarce. By harnessing the power of machine learning, healthcare systems can transition towards a more proactive approach, emphasizing prevention and early intervention over traditional reactive care.

3.9.2 Potential Biases and Limitations

Despite promising results, several limitations must be acknowledged. This study relied solely on the Pima Indians Diabetes Database, which comprises data from female patients of Pima Indian descent. While this dataset is well-known and frequently used in diabetes research, it may not adequately represent the diversity present within the general population. The lack of diversity could introduce biases, affecting the generalizability of the findings to other demographic groups with varying genetic, environmental, or lifestyle characteristics. For instance, differences in diabetes risk factors, such as obesity and physical inactivity, across ethnicities could impact the models' performance.

Additionally, the relatively small size of the dataset (768 records) may limit the models' capacity to capture complex patterns and relationships. A larger, more diverse dataset could offer a broader range of data points, enhancing the models' ability to generalize and improving their robustness when applied to different populations. Expanding the dataset to include more diverse and numerous samples would likely yield more comprehensive insights and enhance model performance across various contexts.

3.9.3 Future Research Directions

To improve generalizability, future research could explore alternative approaches for managing class imbalance beyond SMOTE, such as cost-sensitive learning, which may be more effective for datasets representing diverse population demographics. Additionally, combining ensemble techniques with deep learning approaches could enhance predictive performance, especially on larger, more heterogeneous datasets. Expanding the model’s application to varied populations would help assess the scalability and applicability of ensemble methods for diabetes risk prediction across diverse clinical settings.

4. Conclusion

This study highlights the effectiveness of ensemble learning techniques, particularly random forest and boosting, in predicting diabetes onset with high accuracy and robustness. By effectively managing class imbalance and capturing complex data relationships, these models provide valuable tools for early diabetes diagnosis and management. However, the study emphasizes the need for further research to overcome limitations in dataset diversity and improve model generalizability. Integrating these advanced machine learning models into clinical practice could transform diabetes care by enabling earlier and more precise interventions, improving patient outcomes, and reducing healthcare costs.

5. Acknowledgements

The authors extend their gratitude to the Pima Indians Diabetes Database for supplying the dataset utilized in this study.

6. References

International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). International Diabetes Federation.

American Diabetes Association. (2020). Standards of medical care in diabetes—2020. *Clinical Diabetes, 38*. https://doi.org/10.2337/cd20-as01

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & Wild, S. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results

from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice, 157,* 107843. https://doi.org/10.1016/j.diabres.2019.107843

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice, 138*, 271–281. https://doi.org/10.1016/j.diabres.2018.02.023

Fawwad, A., Govender, D., Ahmedani, M. Y., Basit, A., Lane, J. A., Mack, S. J., & Chandrasekaran, S. (2019). Clinical features, biochemistry, and HLA-DRB1 status in youth-onset type 1 diabetes in Pakistan. *Diabetes Research and Clinical Practice, 149*, 9–17. https://doi.org/10.1016/j.diabres.2019.01.023

Vettoretti, M., Cappon, G., Acciaroli, G., Facchinetti, A., & Sparacino, G. (2018). Continuous glucose monitoring: Current use in diabetes management and possible future applications. *Journal of Diabetes Science and Technology, 12*(5), 1064–1071. https://doi.org/10.1177/1932296818774078

Williams, R., Karuranga, S., Malanda, B., Saeedi, P., Basit, A., Besançon, S., Bommer, C., Esteghamati, A., Ogurtsova, K., Zhang, P., & Colagiuri, S. (2021). Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas, 10th edition. *Diabetes Research and Clinical Practice, 178*, 108072. https://doi.org/10.1016/j.diabres.2021.108072

Nathan, D. M., Buse, J. B., Davidson, M. B., Ferrannini, E., Holman, R. R., & Sherwin, R. (2009). Medical management of hyperglycemia in type 2 diabetes: A consensus algorithm for the initiation and adjustment of therapy. *Diabetes Care, 32*(1), 193–203. https://doi.org/10.2337/dc08-9025

Tankeu, A. T., Noubiap, J. J., Ama Moor, V. J., & Nansseu, J. R. (2021). Epidemiology of diabetes in Africa: Current status, challenges, and perspectives—A systematic review and meta-analysis. *Diabetes Research and Clinical Practice, 175*, 108727. https://doi.org/10.1016/j.diabres.2021.108727

Chen, L., Zhang, X., Li, W., Zhang, Y., & Wang, C. (2022). Machine learning in diabetes prediction and management: Challenges and future perspectives. *Computers in Biology and Medicine, 144*, 105375. https://doi.org/10.1016/j.compbiomed.2022.105375

Mayer, S., Zhang, F. F., & Di, X. (2020). Applications of artificial intelligence in diabetes prediction and management: Emerging trends and future directions. *Journal of Diabetes Science and Technology, 14*(5), 908–916. https://doi.org/10.1177/1932296820917344

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

Meigs, J. B., Wilson, P. W. F., Fox, C. S., Vasan, R. S., Nathan, D. M., & Sullivan, L. M. (2006). Body mass index, metabolic syndrome, and risk of type 2 diabetes or cardiovascular disease. *The Journal of Clinical Endocrinology & Metabolism, 91*(8), 2906–2912. https://doi.org/10.1210/jc.2006-0594

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications, and research directions. *SN Computer Science, 2*, 160. https://doi.org/10.1007/s42979-021-00592-x

Elhaj, H., Achour, N., Tania, M. H., & Aciksari, K. (2023). A comparative study of supervised machine learning approaches to predict patient triage outcomes in hospital emergency departments. *Array, 17*, 100281. https://doi.org/10.1016/j.array.2023.100281

Rashighi, M., & Harris, J. E. (2017). Predicting the future—Big data, machine learning, and clinical medicine. *Physiology & Behavior, 176*, 139–148. https://doi.org/10.1056/NEJMp1606181.Predicting

Subramanian, A., & Li, X. (2019). A review of machine learning approaches in diabetes prediction and management. *Healthcare Analytics, 3*, 100017. https://doi.org/10.1016/j.health.2019.100017

Patel, H., Shah, M., & Patel, P. (2021). Machine learning approaches for predictive modeling of diabetes: A review. *Artificial Intelligence in Medicine, 114*, 102041. https://doi.org/10.1016/j.artmed.2021.102041

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. CRC Press. https://doi.org/10.1201/9781315139470

Smith, J., Roberts, K., Johnson, L., & Brown, M. (2022). Advancements in machine learning applications for chronic disease prediction: A focus on diabetes. *Journal of Biomedical Informatics, 128*, 104042. https://doi.org/10.1016/j.jbi.2022.104042

Zhang, Y., Chen, H., Wang, J., & Li, S. (2023). Deep learning approaches for diabetes prediction: Current progress and future perspectives. *Computers in Biology and Medicine, 149*, 106019. https://doi.org/10.1016/j.compbiomed.2023.106019

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Wu, X., Zhang, Y., Feng, Y., & Xu, L. (2020). Machine learning applications in diabetes prediction and diagnostics. *Journal of Healthcare Engineering, 2020*, 8858435. https://doi.org/10.1155/2020/8858435

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H., & Zhang, Y. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics, 9*, 515. https://doi.org/10.3389/fgene.2018.00515

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data, 5*(1), 42.

https://doi.org/10.1186/s40537-018-0151-6

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Raja, A., Patel, S., & Kannan, P. (2015). Performance analysis of ensemble methods in diabetes prediction. *International Journal of Medical Informatics, 84*(10), 675–682. https://doi.org/10.1016/j.ijmedinf.2015.05.004

Ganie, A. H., Shah, R. A., & Lone, S. M. (2023). Boosting algorithms for improving prediction accuracy in diabetes diagnosis: A comparative study. *Journal of Healthcare Engineering, 2023*, 9937623. https://doi.org/10.1155/2023/9937623

Li, X., Chen, L., & Wang, J. (2024). Optimizing stacking ensemble models with genetic algorithms for diabetes prediction. *Artificial Intelligence in Medicine, 138*, 102584.