

Predictive modeling of early diabetes diagnosis: An evaluation of XGBoost, support vector machine, and random forest classifiers

Idehen Emmanuel Imafidon ^{1,*}, Chikere Obinna Munachiso ², Dominic Evans Onyebuchi ², Adeyemo Latifat Abiodun ³ and Rapuruchukwu Njoku ⁴

¹ Department of Computer Science, Faculty of Computing, National Open University of Nigeria.

² Department of Computer Science, College of Physical and Applied Science, Michael Okpara University of Agriculture, Nigeria.

³ Department of Computer Science, College of Science and Engineering, Osun State University, Nigeria.

⁴ Department of Mechanical Engineering, Faculty of Engineering, Imo State University, Nigeria.

International Journal of Science and Research Archive, 2026, 20(01), 865–878

Publication history: Received on 22 May 2026; revised on 22 July 2026; accepted on 25 July 2026

Article DOI: <https://doi.org/10.30574/ijrsra.2026.20.1.1536>

Abstract

This study addresses the challenge of delayed diagnosis of diabetes, a condition that often leads to severe complications if not detected early. The primary objective is to evaluate and compare the performance of three machine learning classifiers XGBoost, Support Vector Machine (SVM), and Random Forest for early diabetes prediction using clinical and lifestyle data. The study utilizes the Diabetes Health Indicators dataset, which includes features such as body mass index (BMI), blood pressure, cholesterol levels, and physical activity. The dataset was sourced from a publicly available repository and preprocessed through handling missing values, feature scaling, and encoding categorical variables. The models were trained on the processed dataset and evaluated using accuracy, precision, recall, and F1-score metrics, alongside exploratory data analysis to understand feature relationships. Results show that all three models performed effectively, with XGBoost achieving the highest accuracy of 85.11%, followed by SVM at 84.82%, and Random Forest at 83.16%. These findings highlight the strength of ensemble and boosting techniques in handling complex health data and accurately predicting diabetes risk. In conclusion, machine learning models demonstrate strong potential for supporting early diabetes diagnosis and improving clinical decision-making. It is recommended that healthcare systems adopt XGBoost-based predictive models in clinical decision support tools for early screening, while future studies should validate these models using real-world clinical data to enhance reliability and generalizability.

Keywords: Diabetes Diagnosis; Machine Learning; XGBoost; Support Vector Machine; Random Forest; Predictive Modeling.

1. Introduction

The difficulty of detecting diabetes at its early stages has encouraged researchers to explore machine learning as a tool for enhancing diagnostic precision. Among the notable methods is the Extreme Gradient Boosting (XGBoost) algorithm, which has shown strong results when integrated with feature selection strategies such as those provided by Random Forest. Gündoğdu (2023) [1] examined this integration and demonstrated that an XGBoost model supported by Random Forest-based feature selection can accurately identify early diabetes cases. The findings emphasize that this hybrid technique not only strengthens predictive accuracy but also simplifies the model by concentrating on the most relevant variables, thereby contributing to more effective and timely interventions in diabetes management.

* Corresponding author: Idehen Emmanuel Imafidon

Comparative evaluations of machine learning algorithms in medical diagnosis also offer important perspectives on their relative effectiveness across various conditions. Huang et al. (2020) [2], for instance, compared multiple models including Linear Regression, Random Forest, Support Vector Regression, XGBoost, LASSO Regression, and ensemble techniques in predicting blood pressure during hemodialysis, a task that shares methodological similarities with diabetes prediction. Their results revealed that while individual models have distinct advantages, ensemble strategies frequently surpass single algorithms because they integrate outputs from different approaches. This suggests that for early diabetes detection, combining models such as SVM, Random Forest, and XGBoost may provide superior outcomes by leveraging their complementary strengths in handling diverse data features.

Research on diabetes diagnosis using machine learning has consistently shown that these methods can uncover intricate relationships in health-related datasets. Farajollahi et al. (2021) [3] highlighted this in their study, illustrating that algorithms such as Support Vector Machines and Random Forests are particularly effective when carefully tuned and tailored to the specifics of medical data. Their findings stress the critical role of feature selection and parameter optimization in boosting diagnostic accuracy. This supports the broader argument that an in-depth understanding of how each model interacts with clinical data is vital to developing dependable predictive tools for early diabetes screening.

Efforts to design models that are both accurate and interpretable have also gained attention. Hao et al. (2022) [4] introduced novel techniques for extracting decision rules from biased Random Forests and Fuzzy Support Vector Machines, aimed at making early diabetes diagnosis more transparent. Their study underscores the necessity of interpretability in medical AI, as clinical practitioners must be able to trust and understand the rationale behind predictions. By providing explicit rules, these methods help bridge the gap between predictive performance and clinical trust, thereby encouraging wider acceptance of machine learning in healthcare. This line of research enhances the strengths of algorithms such as XGBoost and SVM while making their decision-making processes more understandable and practical in real-world clinical applications.

Diabetes has become an escalating public health issue worldwide, with its prevalence rising steadily due to lifestyle transitions, reduced physical activity, and unhealthy dietary patterns. Detecting the disease at an early stage is vital to minimizing complications, improving health outcomes, and lowering healthcare expenditures. Conventional diagnostic approaches, such as fasting blood sugar tests and insulin measurements, remain useful but are often costly, time-intensive, and typically identify diabetes only after complications have already developed. In recent years, machine learning (ML) methods have gained attention as powerful alternatives for early prediction and diagnosis. These algorithms are capable of processing extensive health datasets and uncovering subtle patterns that might go unnoticed with traditional diagnostic techniques by integrating features such as body mass index (BMI), age, and family history (Modak & Jha, 2024) [5].

Among the numerous ML algorithms, Extreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), and Random Forests (RF) stand out for their strong predictive ability in diabetes-related studies. XGBoost, a gradient boosting technique, is widely recognized for its excellent accuracy and its ability to manage missing values while capturing interactions between features. Its capacity to model complex associations between variables makes it especially well suited for predictive tasks in healthcare [6]. SVMs, which classify data by constructing an optimal hyperplane to separate classes, have also been frequently applied in medical diagnostics. Their effectiveness in dealing with high-dimensional data renders them particularly advantageous when working with large feature sets, as seen in diabetes prediction models [7]. Meanwhile, Random Forest, an ensemble learning approach, is highly regarded for its robustness, its ability to reduce overfitting, and its reliable predictions produced through combining multiple decision trees. Such strengths make Random Forest highly valuable in clinical contexts, where variability and noise in patient data are common [8].

Several investigations have compared the effectiveness of these algorithms in diabetes prediction, often showing that ensemble methods such as Random Forest and XGBoost yield better performance than individual classifiers. Modak and Jha (2024) [5] reported that both XGBoost and RF achieved impressive accuracy in predicting diabetes by incorporating a combination of clinical measures and lifestyle indicators. Their results suggest that these models provide faster and more efficient diagnostic support compared with traditional techniques. Similarly, Ghosh et al. (2024) [8] assessed multiple machine learning models and concluded that ensemble-based approaches offered greater predictive accuracy and generalization capabilities. Their work also emphasized the significance of careful preprocessing and feature selection, since irrelevant or noisy attributes can negatively influence performance. Together, these findings reinforce the growing evidence that machine learning tools can enhance early diagnosis, equipping clinicians with more reliable support systems and enabling patients to receive earlier interventions.

Nevertheless, implementing ML algorithms in diabetes diagnosis is not without limitations. Challenges include ensuring high-quality, representative datasets, addressing interpretability concerns, and minimizing overfitting. For clinical adoption, transparency of model decision-making is essential. Rule extraction techniques used in methods such as SVM [7] and the interpretability of decision trees within Random Forests offer potential solutions by making predictions more understandable to healthcare providers. In addition, feature selection strategies such as those demonstrated by Xu and Wang (2019) [6] in their risk prediction model are crucial for improving accuracy and computational efficiency. Addressing these barriers can unlock the full potential of ML-driven models in diabetes diagnosis, ultimately supporting earlier detection, personalized treatment, and improved patient outcomes.

Detecting diabetes early is essential for effective disease management and for preventing such complications. Unfortunately, conventional diagnostic techniques, including insulin testing and blood glucose measurement, are often costly, time-intensive, and tend to identify the illness only after significant damage has already occurred. This delay in detection highlights the need for affordable, efficient, and timely diagnostic methods for early-stage prediction. While conventional clinical tools remain valuable, they are often inadequate in terms of speed and accuracy, particularly in low-resource healthcare environments. Machine learning (ML), with its capacity to analyze massive datasets and reveal hidden patterns, presents a strong opportunity to bridge this diagnostic gap (Alzboon et al., 2023) [9].

One of the central difficulties in early diabetes detection lies in the complexity and variability of the disease. Multiple factors including genetics, age, lifestyle behaviors, and medical history interact to influence risk, making traditional prediction techniques insufficient. Machine learning models such as Support Vector Machines (SVM), Random Forest (RF), and advanced ensemble algorithms like XGBoost have demonstrated encouraging results in handling large-scale, multidimensional datasets to detect relevant predictors of diabetes. These models can integrate diverse data sources, ranging from demographic and clinical measures to lifestyle indicators, thereby enhancing predictive precision. However, their performance can be undermined by issues such as overfitting, poor feature selection, or the use of low-quality and unrepresentative datasets. Consequently, despite the increasing research on ML for diabetes diagnosis, the most effective approach for early prediction remains uncertain, and further research is needed to establish reliable strategies for model design, validation, and real-world application (Choudhary et al., 2022; Tahsin et al., 2022) [10,11].

Another important obstacle is ensuring that machine learning models are both interpretable and clinically applicable. Although algorithms like Random Forest and XGBoost deliver strong predictive performance, their inherent “black-box” nature often limits their adoption in healthcare, where transparency and trust are vital. Clinicians require not only accurate predictions but also clear justifications that can guide medical decisions with confidence. Additionally, many ML models demand high computational power, which poses a challenge in settings with limited technological resources.

Therefore, the challenge extends beyond developing accurate prediction systems to creating models that are explainable, efficient, and widely accessible. Current research is focusing on hybrid approaches, feature selection methods, and rule extraction techniques to improve transparency while reducing computational costs without sacrificing predictive power (Albahli, 2020; Alzboon et al., 2023) [12,9]. Ensuring that such models are practical, scalable, and reliable for early diabetes detection is essential for their successful adoption in healthcare practice.

Aim and Objectives of the Study

The aim of this study is to evaluate the performance of XGBoost, Support Vector Machine (SVM), and Random Forest classifiers for the early diagnosis of diabetes.

The objectives are to:

- Review existing studies on machine learning models for diabetes prediction, focusing on the effectiveness of XGBoost, SVM, and Random Forest in medical diagnostics.
- Develop a predictive model for diabetes by integrating relevant clinical and lifestyle features and performing comprehensive data preprocessing to prepare the dataset for accurate model training.
- Implement XGBoost, SVM, and Random Forest classifiers using the prepared dataset, followed by model training and testing.
- Evaluate the models based on accuracy, precision, recall, and F1-score, and compare their performance to identify the most effective method for early diabetes diagnosis.

2. Materials and Methods

2.1. Overview

This section provides a comprehensive overview of the processes involved in developing and evaluating machine learning models for early diabetes diagnosis. It begins by describing the data collection process, including the dataset used, its features, and the preprocessing techniques applied (e.g., normalization, missing value imputation). The chapter details the specific machine learning algorithms used XGBoost, Support Vector Machine (SVM), and Random Forest explaining their theoretical foundations and why they were selected for this study. Additionally, the chapter outlines the process of model training, cross-validation, and hyperparameter tuning to optimize the performance of each classifier.

2.2. Dataset Description

The Diabetes Health Indicators Dataset (diabetes_012_health_indicators_BRFSS2015) available on Kaggle is a detailed collection of health-related information focusing on individuals at risk of diabetes. It originates from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), a nationwide survey that gathers data on health behaviors, chronic conditions, and preventive health practices. The dataset encompasses numerous features, including demographic characteristics (such as age, gender, and race), lifestyle factors (like physical activity, smoking habits, and alcohol use), and health indicators (including body mass index, blood pressure, and cholesterol levels). This extensive range of attributes makes the dataset highly valuable for examining patterns and predicting diabetes risk among individuals.

The dataset contains 101,766 records with 21 features, comprising both categorical and continuous variables. These features capture important health metrics, including BMI, blood pressure, cholesterol levels, and history of diabetes-related conditions. Its key target variable is binary, indicating whether an individual has been diagnosed with diabetes (1) or not (0), making it particularly suitable for classification-based predictive modeling. The dataset's size and feature diversity provide a strong foundation for applying machine learning algorithms to analyze risk factors, support early diagnosis, and build predictive models using classifiers such as XGBoost, Support Vector Machines, and Random Forest. The dataset has a file size of 21.6 MB. Table 1 presents a detailed description of the dataset.

Table 1 Data Description

Column Name	Description	Example Values
Diabetes_012	Indicates if the individual has been diagnosed with diabetes (1 = Yes, 0 = No).	0, 1
HighBP	Whether the individual has high blood pressure (1 = Yes, 0 = No).	0, 1
HighChol	Whether the individual has high cholesterol (1 = Yes, 0 = No).	0, 1
CholCheck	Whether the individual had a cholesterol check in the past 5 years (1 = Yes, 0 = No).	0, 1
BMI	Body Mass Index, a measure of body fat based on height and weight.	22.5, 30.1
Smoker	Whether the individual is a smoker (1 = Yes, 0 = No).	0, 1
Stroke	Whether the individual has ever experienced a stroke (1 = Yes, 0 = No).	0, 1
HeartDiseaseorAttack	Whether the individual has had heart disease or a heart attack (1 = Yes, 0 = No).	0, 1
PhysActivity	Whether the individual engages in regular physical activity (1 = Yes, 0 = No).	0, 1
Fruits	Whether the individual consumes fruits regularly (1 = Yes, 0 = No).	0, 1
Veggies	Whether the individual consumes vegetables regularly (1 = Yes, 0 = No).	0, 1
HvyAlcoholConsump	Whether the individual consumes alcohol heavily (1 = Yes, 0 = No).	0, 1
AnyHealthcare	Whether the individual has access to healthcare (1 = Yes, 0 = No).	0, 1

NoDocbcCost	Whether the individual skipped doctor visits due to cost (1 = Yes, 0 = No).	0, 1
GenHlth	General health status (1 = Excellent, 5 = Poor).	1, 2, 3, 4, 5
MentHlth	Number of days in the past 30 when mental health was not good.	0, 5, 15
PhysHlth	Number of days in the past 30 when physical health was not good.	0, 3, 10
DiffWalk	Whether the individual has difficulty walking or climbing stairs (1 = Yes, 0 = No).	0, 1
Sex	Sex of the individual (1 = Male, 2 = Female).	1, 2
Income	Household income level, reported on an ordinal scale from lowest to highest income bracket.	1, 4, 8
Age	Age of the individual.	25, 58, 67
Education	Highest education level attained (1 = Less than high school, 5 = College).	1, 2, 3, 4, 5

2.2.1. Justification of Using XGBoost (Extreme Gradient Boosting)

XGBoost is chosen due to its exceptional performance on large and complex datasets such as the Diabetes Health Indicators dataset. This dataset contains a mix of numerical and categorical variables, and XGBoost efficiently manages these heterogeneous data types. Built on gradient boosting, XGBoost iteratively refines models by correcting the errors of prior iterations, allowing it to capture subtle patterns that are crucial for early diabetes prediction. Its robustness in handling missing values, outliers, and noisy data makes it particularly suitable for classification tasks. Furthermore, XGBoost incorporates regularization techniques to minimize overfitting, ensuring the model generalizes well to unseen data a key requirement for real-world applications. Additionally, XGBoost provides feature importance metrics, which can highlight the most influential health indicators, enhancing the interpretability of the predictive model.

2.2.2. Justification of Using Support Vector Machine (SVM)

Support Vector Machine (SVM) is well-suited for this dataset because of its capacity to handle high-dimensional data and complex, non-linear relationships between features. The Diabetes Health Indicators dataset contains numerous health-related variables, some of which may interact in intricate ways with the target variable. SVM maps these features into a higher-dimensional space using kernel functions to identify an optimal separating hyperplane, maximizing the margin between diabetic and non-diabetic classes. This allows the model to uncover subtle patterns in health indicators, which is critical for accurate diabetes prediction. SVMs are also resilient against overfitting, especially with proper hyperparameter tuning, ensuring reliable performance even when data is noisy or variable. This robustness makes SVM a dependable choice for early disease detection.

2.2.3. Justification of Using Random Forest (RF)

Random Forest is selected for its ensemble approach, combining multiple decision trees to produce more stable and accurate predictions. Given the richness of features in the Diabetes Health Indicators dataset, Random Forest effectively handles large datasets with multiple predictors. By training each tree on a random subset of the data, the algorithm reduces overfitting and improves generalization. Its ability to capture complex, non-linear interactions between variables is particularly valuable in healthcare applications, where such interactions are common. Random Forest also provides insights into feature importance, helping to identify the key health indicators influencing diabetes risk. Furthermore, its seamless handling of both categorical and numerical data makes it a robust and interpretable choice for early diabetes prediction.

2.3. Implementation Environment

Hardware Requirement: A multi-core processor (e.g., Intel i7 or AMD Ryzen 7) is required to efficiently handle computations and speed up the training of machine learning models like XGBoost, SVM, and Random Forest. At least 16 GB of RAM is recommended to support data processing and ensure smooth model training without performance degradation. A minimum of 500 GB of storage, preferably an SSD, is needed to store datasets and model files, ensuring fast data access and processing. A dedicated GPU (e.g., NVIDIA GTX/RTX) can accelerate computations, especially for larger datasets or complex models, but is not strictly required for this system. A stable internet connection is needed for downloading datasets and libraries and for cloud-based operations, if necessary.

Software Requirement: The implemented system is compatible with Windows 10 or later, macOS, or Linux distributions (such as Ubuntu). Python (3.x or later) is the primary programming language used for implementing machine learning algorithms such as XGBoost, SVM, and Random Forest, supporting libraries required for data preprocessing, model training, and evaluation, including Scikit-learn, XGBoost, Pandas, NumPy, and Matplotlib/Seaborn. Jupyter Notebook, Google Colab, PyCharm, or Visual Studio Code serve as suitable Integrated Development Environments (IDEs) for the interactive analysis and model development performed in this study.

2.4. System Design

The implemented system for early diabetes detection is designed to employ machine learning models to classify individuals according to their risk of developing diabetes, providing a tool that can assist both healthcare providers and patients in disease prevention and management. It utilizes the Diabetes Health Indicators dataset, which incorporates critical variables such as BMI, blood pressure, cholesterol, and lifestyle factors like smoking and physical activity. By implementing advanced algorithms such as Random Forest (RF), Support Vector Machine (SVM), and XGBoost, the system predicts whether an individual is at risk of diabetes. These algorithms are chosen for their ability to handle the intricacies of medical data, including capturing non-linear patterns and managing large-scale datasets efficiently. The dataset undergoes preprocessing steps, including handling missing values, feature scaling, and encoding categorical variables, before training the models. Model performance is then assessed using evaluation metrics such as accuracy, precision, recall, and F1-score. The ultimate aim of the system is to enable early intervention through reliable predictions, promoting data-driven decision-making in diabetes management and contributing to the reduction of disease prevalence by facilitating timely detection.

The system is built from several key components that ensure effective data handling and reliable model training. The first is the Data Preprocessing Module, which prepares the raw dataset for modeling by cleaning and organizing the data. This involves addressing missing or redundant entries, managing outliers, normalizing numerical variables, and encoding categorical data into formats compatible with machine learning algorithms. Following this, the Model Training and Evaluation Module applies the selected algorithms Random Forest, Support Vector Machine, and XGBoost to the processed dataset. This module not only trains the models but also evaluates their performance using metrics including accuracy, confusion matrix, precision, recall, and F1-score. The Visualization Module provides graphical outputs such as bar charts, ROC curves, and confusion matrices, which help users interpret model results effectively. These visualizations enable a clear comparison of each model's performance and guide decisions regarding which model is most suitable for deployment. Although the system currently lacks a user interface, the design allows for future development to enable healthcare professionals or patients to input data and obtain predictions. Lastly, the Reporting and Result Interpretation Module analyzes evaluation outcomes, offering insights into each model's performance and recommending the most appropriate model based on criteria established for early diabetes detection.

The system workflow is organized to ensure that each step is carried out in a logical and efficient sequence. The process begins with loading the dataset and executing the Data Preprocessing Module, where the data is cleaned and transformed. Missing values are either imputed or removed, categorical variables are encoded using one-hot or label encoding techniques, and numerical features are scaled to maintain consistency across all models. Once preprocessing is complete, the data is divided into training and testing subsets, typically using an 80:20 split. Next, the system implements the three selected machine learning algorithms: Random Forest, Support Vector Machine, and XGBoost. These models are trained on the training set to learn the patterns and correlations between health indicators and diabetes risk. Following training, the models are evaluated on the testing set using metrics such as accuracy, precision, recall, and F1-score to measure their effectiveness in correctly classifying individuals. The Visualization Module then generates comparative charts, including confusion matrices and ROC curves, providing a clear graphical view of model performance and highlighting their respective strengths and weaknesses. Finally, the model demonstrating the highest accuracy and overall performance is chosen as the optimal candidate for deployment. This workflow ensures that the system delivers reliable and interpretable results, aiding healthcare professionals in early diabetes detection and intervention.

The expected outcome of the system is a detailed comparison of the predictive performance of Random Forest, Support Vector Machine, and XGBoost for early diabetes detection. Using multiple evaluation metrics, the system identifies the model that performs best for this task. Analysis of confusion matrices and ROC curves illustrates each model's ability to classify individuals as diabetic or non-diabetic, highlighting key measures such as precision, recall, and rates of false positives and negatives. The system also assesses which model generalizes most effectively to new, unseen data, ensuring that the selected model is robust and not overfitted. Figure 1 illustrates the working methodology of the implemented system.

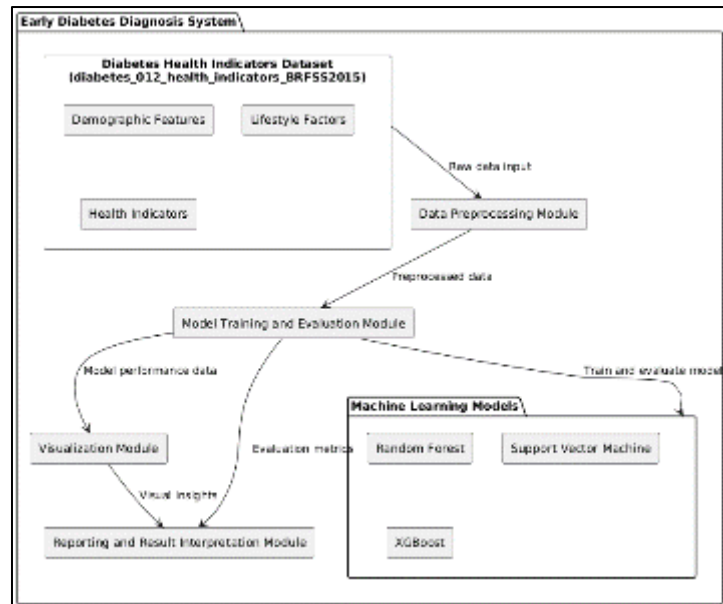


Figure 1 Working Methodology of the Implemented System

2.4.1. Algorithm of the Implemented Model

Input Data: Diabetes Health Indicators Dataset.

(i) Load Data: Import the Diabetes Health Indicators dataset. (ii) Preprocess Data: Clean, normalize, and encode the data. (iii) Split Data: Divide data into training and testing sets. (iv) Train Models: Train models using Random Forest, SVM, and XGBoost. (v) Evaluate Models: Evaluate models with accuracy, precision, recall, and F1-score. (vi) Visualize Results: Generate ROC curves and confusion matrices. (vii) Interpret Results: Select the best-performing model. (viii) Deploy Model: Deploy the chosen model for prediction.

Implemented System Flowchart

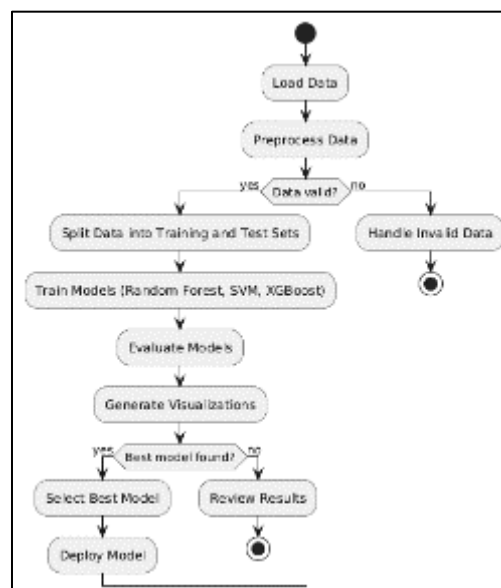


Figure 2 Implemented System Flowchart

2.4.2. Selection of the Input and Output Data

The diabetes prediction system utilizes a dataset comprising multiple health indicators that are essential for evaluating an individual's risk of developing diabetes. The input features chosen for the model are comprehensive, covering a range of health, lifestyle, and demographic variables. These features include Body Mass Index (BMI), Blood Pressure, Cholesterol Levels, Physical Activity (PhysActivity), Smoking status (Smoker), and the presence of health conditions such as Stroke and Heart Disease, along with demographic factors like Age, Sex, Income, and Education. Collectively, these features serve as important predictors, integrating both medical history and lifestyle information to assess the likelihood of diabetes. The target variable, Diabetes_012, indicates an individual's diabetes status: "0" for no diabetes, "1" for pre-diabetes, and "2" for confirmed diabetes. For the purposes of classification in this study, this three-class variable was subsequently converted into a binary target, with pre-diabetes and confirmed diabetes combined into a single positive class, and no diabetes retained as the negative class, consistent with the binarization approach described in Section 3.1. This arrangement enables the model to learn from a diverse set of factors that may influence an individual's risk of diabetes.

The choice of input features was made deliberately to provide a thorough perspective of an individual's health and lifestyle, ensuring the prediction model is both robust and reliable. The dataset's mix of numerical and categorical features such as BMI measurements and smoking habits allows for appropriate preprocessing steps, including scaling and encoding. Once missing values are addressed and the data is normalized, the features are partitioned into training and testing sets, with Diabetes_012 serving as the dependent variable for prediction. This configuration guarantees that the model is trained on a rich array of health-related attributes and can produce meaningful outputs, offering insights into an individual's risk of diabetes. Such a setup is valuable for early detection and informed intervention in healthcare decision-making.

3. Results and Discussion

3.1. Data Loading and Initial Processing

The dataset is initially imported from a local CSV file containing health-related metrics used for diabetes prediction. To streamline the classification process, the original target variable is transformed into a binary format, combining individuals with prediabetes or diabetes into a single positive class, while those without diabetes form the negative class. This approach simplifies the model's task by focusing on distinguishing individuals at risk from healthy ones. The balance between these two classes is visualized to check for any skew in the dataset. Furthermore, any missing values are removed to ensure the dataset's integrity and to prevent issues during model training.

To gain a clearer understanding of the relationships between key health indicators and diabetes outcomes, multiple visualizations are generated. Kernel Density Estimate (KDE) plots illustrate the distributions of critical features such as BMI, Age, High Blood Pressure, High Cholesterol, and Physical Activity, highlighting variations between diabetic and non-diabetic individuals. In addition, boxplots are employed to assess the range and central tendencies of numerical variables like BMI, Age, Mental Health, Physical Health, and Income relative to diabetes status. These graphical analyses provide essential insights into the characteristics that distinguish at-risk individuals, guiding both feature selection and model development, as demonstrated in Figures 3 to 8 below.

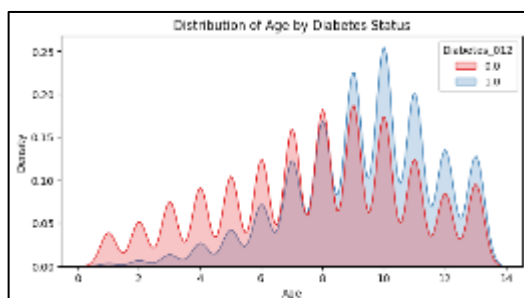


Figure 3 Distribution of Age by Diabetes Status

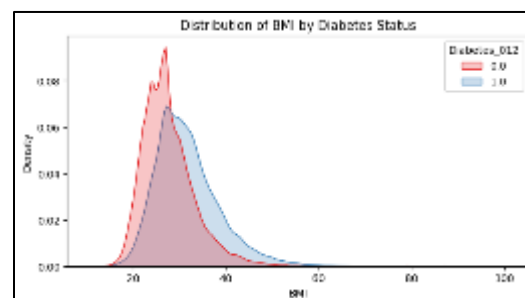


Figure 4 Distribution of BMI by Diabetes Status

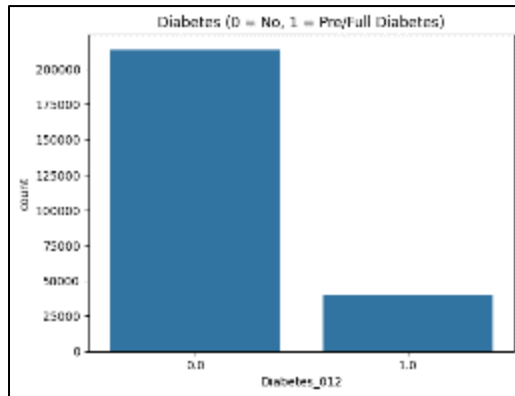


Figure 5 Diabetes vs Count

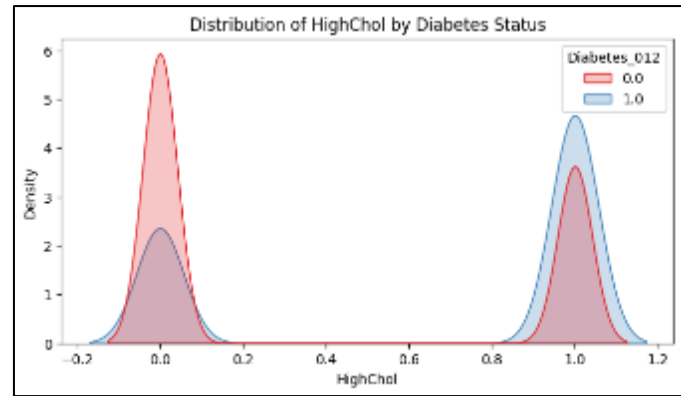


Figure 6 Distribution of HighChol by Diabetes Status

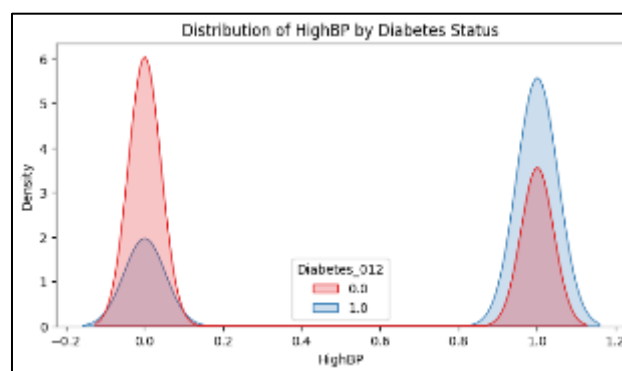


Figure 7 Distribution of HighBP by Diabetes Status

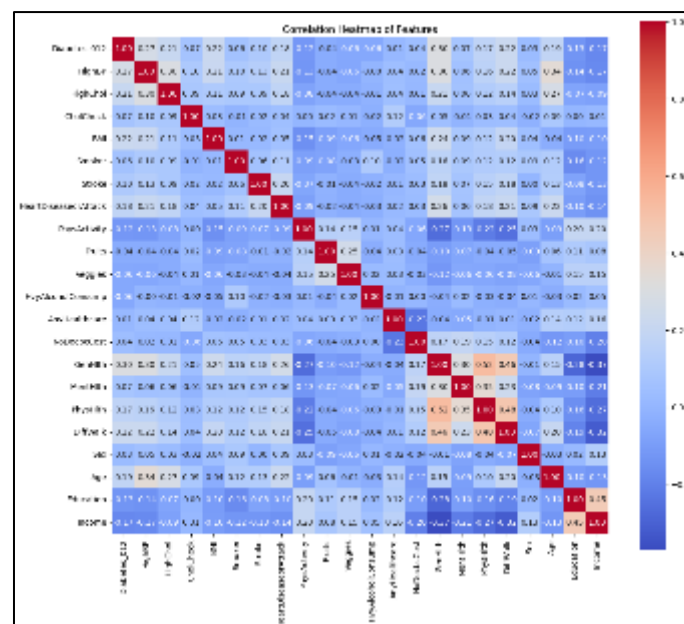


Figure 8 Correlation Heatmap of Features

3.2. Evaluation: Random Forest

The Random Forest model was applied to classify individuals according to their diabetes risk and achieved a total accuracy of 83.16%, as summarized in Table 2. The classification report indicates strong performance for the model, demonstrating high precision (0.87) and recall (0.94), which together yield a robust F1-score of 0.90 for the positive

class. The confusion matrix, displayed in Figure 9, provides a detailed breakdown of true positives, false negatives, true negatives, and false positives, offering a clear view of the model's prediction behavior. Additionally, Figure 10 presents the Feature Importance, highlighting which health indicators contributed most significantly to the prediction outcomes.

Table 2 Random Forest Classification Report

Class/Metric	Precision	Recall	F1-Score	Support
0.0	0.87	0.94	0.90	42795
1.0	0.43	0.25	0.32	7941
Accuracy			0.83	50736
Macro Avg	0.65	0.60	0.61	50736
Weighted Avg	0.82	0.83	0.81	50736

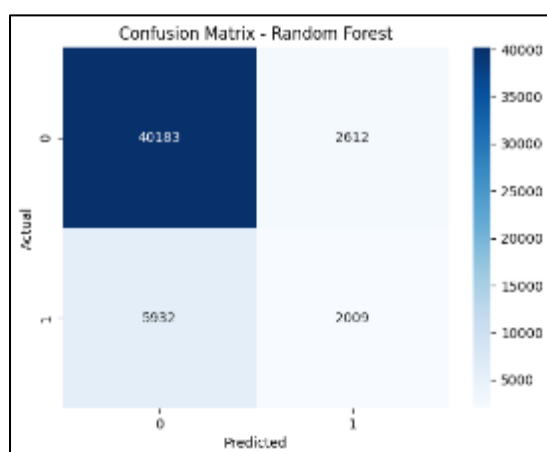


Figure 9 Random Forest Confusion Matrix

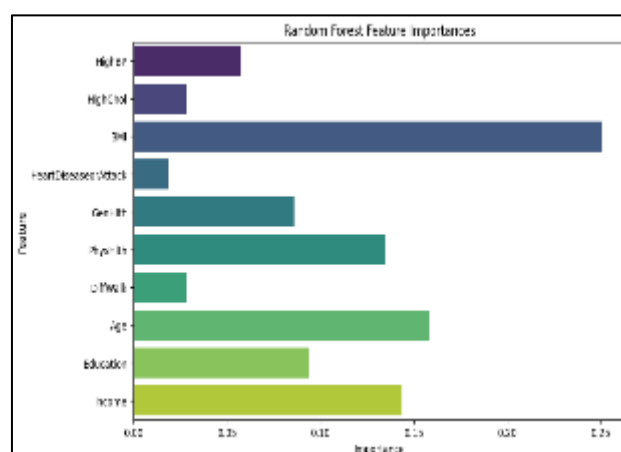


Figure 10 Random Forest Feature Importance

3.3. Evaluation: XGBoost

The XGBoost model demonstrated a strong performance in predicting diabetes risk, achieving an overall accuracy of 85.11%. It shows excellent classification capabilities, with a precision of 0.87 and a notably high recall of 0.97, which combine to produce a robust F1-score of 0.92 for the positive class, as summarized in Table 3. The confusion matrix, displayed in Figure 11, provides a detailed account of true positives, false negatives, true negatives, and false positives, offering insight into the model's prediction accuracy. Furthermore, Figure 12 illustrates the Feature Importance, highlighting which variables were most influential in determining diabetes risk.

Table 3 XGBoost Classification Report

Class/Metric	Precision	Recall	F1-Score	Support
0.0	0.87	0.97	0.92	42795
1.0	0.57	0.19	0.29	7941
Accuracy			0.85	50736
Macro Avg	0.72	0.58	0.60	50736
Weighted Avg	0.82	0.85	0.82	50736

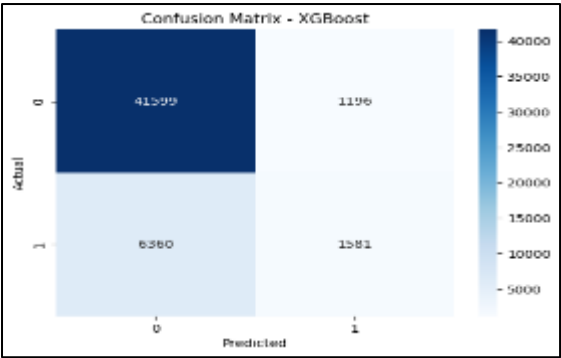


Figure 11 XGBoost Confusion Matrix

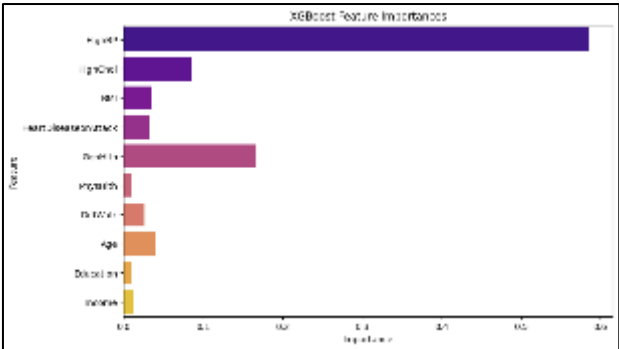


Figure 12 XGBoost Feature Importance

3.4. Evaluation: SVM

The Support Vector Machine (SVM) model achieved a notable overall accuracy of 84.82% in predicting diabetes risk. It exhibited strong classification performance, with a precision of 0.86 and an exceptionally high recall of 0.99, resulting in an F1-score of 0.92 for the positive class, as summarized in Table 4. The confusion matrix, depicted in Figure 13, provides a breakdown of true positives, false negatives, true negatives, and false positives, demonstrating the model's effectiveness in accurately identifying individuals at risk for diabetes.

Table 4 SVM Classification Report

Class/Metric	Precision	Recall	F1-Score	Support
0.0	0.86	0.99	0.92	42795
1.0	0.59	0.10	0.17	7941
Accuracy			0.85	50736
Macro Avg	0.72	0.54	0.54	50736
Weighted Avg	0.81	0.85	0.80	50736

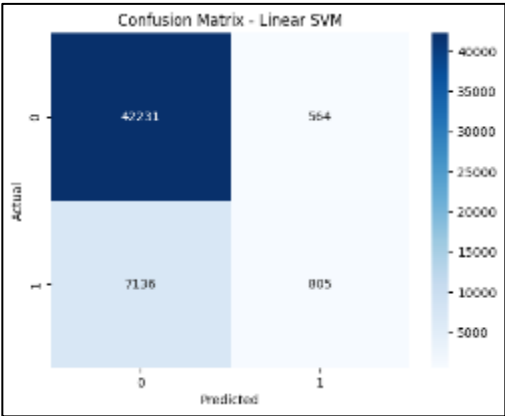


Figure 13 SVM Confusion Matrix

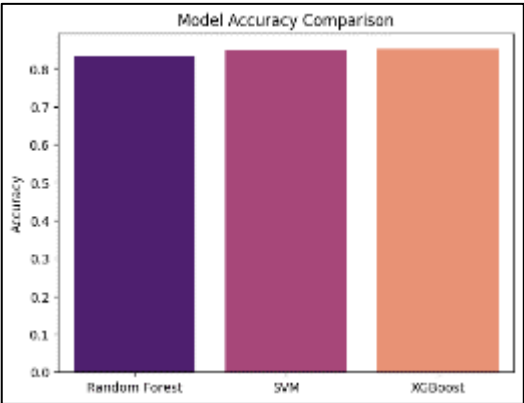


Figure 14 Model Accuracy Comparison

3.5. Model Comparison

XGBoost achieved the highest accuracy at 85.11%, followed by SVM with 84.82%, and Random Forest at 83.16%. All models demonstrate strong overall performance, with XGBoost showing the best ability to generalize to unseen data, as shown in Figure 14 and Table 5.

Table 5 Model Accuracy Comparison

Model	Accuracy
XGBoost	85.11%
SVM	84.82%
Random Forest	83.16%

4. Discussion

This study evaluated the design and development of a predictive system aimed at early diabetes detection using three machine learning techniques: Random Forest, Support Vector Machine, and XGBoost. The Diabetes Health Indicators dataset underwent thorough preprocessing, including cleaning, encoding of categorical features, and scaling, to optimize it for modeling. Exploratory visualizations were conducted to examine the distributions of key health metrics in relation to diabetes status. Each algorithm was trained and tested on the processed dataset, with performance evaluated through metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Among the models, XGBoost demonstrated the highest accuracy at 85.11%, followed closely by SVM and Random Forest. Overall, the findings highlight the effectiveness of machine learning methods in facilitating early diabetes prediction and set the stage for further enhancements and practical applications in healthcare environments.

5. Conclusion

This paper highlights the significant potential of machine learning models in facilitating early diabetes detection. By applying XGBoost, SVM, and Random Forest to health-related data, the research showed that predictive analytics can serve as a powerful support tool for clinical decision-making. These models are capable of identifying intricate patterns within the data that may not be obvious through conventional diagnostic methods, enabling earlier recognition of individuals at risk before severe symptoms arise.

Among the algorithms tested, XGBoost demonstrated the strongest performance, indicating its ability to handle heterogeneous and potentially noisy healthcare data while maintaining high predictive accuracy. This underscores the suitability of gradient boosting techniques for medical datasets that involve multiple interrelated variables. Meanwhile, the robust performance of SVM and Random Forest confirms the value of exploring different algorithmic approaches to select the most appropriate model for a given dataset and prediction task.

Although the models showed promising accuracy, variations in their ability to correctly identify positive diabetes cases illustrate a common challenge in medical diagnostics: balancing sensitivity and specificity. The study emphasizes that no single model can perfectly detect every case, particularly for complex diseases like diabetes with multifactorial risk factors. Consequently, combining machine learning insights with clinical judgment remains essential for effective early diagnosis and patient care.

Overall, this research contributes to the growing evidence supporting the integration of artificial intelligence in healthcare. The successful use of these classifiers demonstrates that machine learning can enhance early diabetes detection, potentially improving patient outcomes through timely interventions. Ongoing development and refinement of these models will further strengthen their effectiveness, opening opportunities for precision medicine and personalized healthcare strategies.

Recommendations

To further strengthen the predictive accuracy of diabetes diagnosis models, it is advisable to incorporate a wider array of clinical and lifestyle features. Adding biomarkers such as blood glucose levels, HbA1c, and insulin measurements could provide more nuanced information about an individual's metabolic health. Moreover, including longitudinal data that tracks changes over time may allow the model to detect early trends in diabetes progression, facilitating timely intervention before the condition advances.

Future research should also prioritize optimizing model performance through advanced strategies such as hyperparameter tuning, systematic feature selection, and ensemble methods. The use of automated machine learning (AutoML) frameworks could simplify this process by quickly identifying optimal model configurations. Additionally,

exploring deep learning techniques may offer further enhancements, particularly with larger, more complex datasets, as these architectures are adept at capturing intricate nonlinear relationships among variables.

Another crucial recommendation involves effectively addressing class imbalance, as current models tend to be less sensitive to diabetes-positive cases. Methods like synthetic minority oversampling (SMOTE), cost-sensitive learning, or tailored loss functions can help rebalance the training dataset and improve detection of minority classes. Reducing false negatives is particularly vital in medical contexts, where failing to identify positive cases can have serious health implications.

Finally, for real-world implementation, these models should be integrated into user-friendly platforms and connected with existing electronic health record systems. Engaging healthcare professionals in the design process will help create practical, actionable decision-support tools. Continuous validation using updated patient data, coupled with feedback from clinical deployment, will ensure that the models remain accurate, relevant, and valuable for early diabetes detection over time.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Statement of ethical approval

This study relies exclusively on the publicly available, de-identified Diabetes Health Indicators Dataset (BRFSS 2015), obtained from the Kaggle repository. The present research work does not contain any studies performed on animals or human subjects directly by the authors.

Statement of informed consent

Not applicable. This study did not involve direct interaction with human participants; it used a publicly available, de-identified secondary dataset.

References

- [1] Gündoğdu S. Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimed Tools Appl.* 2023;82(22):34163-81.
- [2] Huang JC, Tsai YC, Wu PY, Lien YH, Chien CY, Kuo CF, et al. Predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method. *Comput Methods Programs Biomed.* 2020;195:105536.
- [3] Farajollahi B, Mehmannaavaz M, Mehrjoo H, Moghbeli F, Sayadi MJ. Diabetes diagnosis using machine learning. *Front Health Inform.* 2021;10(1):65.
- [4] Hao J, Luo S, Pan L. Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes. *Sci Rep.* 2022;12(1):9858.
- [5] Modak SKS, Jha VK. Diabetes prediction model using machine learning techniques. *Multimed Tools Appl.* 2024;83(13):38523-49.
- [6] Xu Z, Wang Z. A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and XGBoost ensemble classifier. In: *Proceedings of the 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI)*; 2019 Jun 7-9; Guilin, China. IEEE; 2019. p. 278-83.
- [7] Singh N, Singh P, Bhagat D. A rule extraction approach from support vector machines for diagnosing hypertension among diabetics. *Expert Syst Appl.* 2019;130:188-205.
- [8] Ghosh S, Acharya R, Jana ND. An experimental analysis of machine learning models for diabetes classification. In: *Proceedings of the International Conference on Communication, Devices and Networking*; 2024 Jan. Singapore: Springer Nature Singapore; p. 125-38.

- [9] Alzboon MS, Al-Batah MS, Alqaraleh M, Abuashour A, Hamadah Bader AF. Early diagnosis of diabetes: a comparison of machine learning methods. *Int J Online Biomed Eng.* 2023;19(15).
- [10] Choudhary S, Kumar A, Choudhary S. Prediction and comparison of diabetes with logistic regression, naïve Bayes, random forest, and support vector machine. In: *Proceedings of the International Conference on Innovations in Computer Science and Engineering*; 2022 Sep. Singapore: Springer Nature Singapore; p. 273-83.
- [11] Tahsin MS, Jobayer M, Antor MBU, Islam M, Raisa FF, Shaikat MAH. Predictive analysis & brief study of early-stage diabetes using multiple classifier models. In: *Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*; 2022 Jan 26-29; Las Vegas, NV, USA. IEEE; 2022. p. 203-7.
- [12] Albahli S. Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection. *J Med Imaging Health Inform.* 2020;10(5):1069-75.