Performance Comparison of Random Forest and XGBoost for Diabetes Prediction

Dr. Vijay Kumar Samyal[1] , Aditya Kumar[2]

[1]Associate Professor, Department of CSE, MIMIT Malout
[2]Student, Department of CSE, MIMIT Malout

Abstract :
Machine learning is widely used in modern healthcare systems to support disease prediction and diagnosis. Diabetes is one of the most common chronic diseases, and early prediction is important to reduce long-term risks. This paper compares two popular machine learning models, Random Forest and XGBoost, for diabetes prediction. A publicly available diabetes dataset is used, and the models are evaluated using accuracy, precision, recall, and F1-score. Experimental results show that XGBoost performs slightly better in accuracy and precision, while Random Forest performs competitively with simpler tuning and faster execution. The study concludes that both models are effective, but XGBoost provides better overall performance for healthcare prediction.

Keywords: Machine Learning, Diabetes Prediction, Random Forest, XGBoost, Healthcare, Classification, Accuracy.

1.INTRODUCTION

Diabetes is one of the fastest-growing chronic diseases across the world. According to medical studies, millions of people are affected each year because of high blood glucose levels, obesity, and lifestyle habits. Early prediction of diabetes can help in treatment planning and can reduce long-term health complications. Machine learning techniques have become very useful in disease prediction because they can analyze complex medical data and discover hidden patterns.

In recent years, many machine learning algorithms have been used for diabetes prediction, such as Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting. However, Random Forest and XGBoost are among the most powerful models because they can handle nonlinear relationships and produce high accuracy. Random Forest uses multiple decision trees, while XGBoost improves learning by boosting weak models and reducing errors.

This research focuses on comparing the performance of Random Forest and XGBoost for predicting diabetes. A publicly available healthcare dataset is used, and the models are evaluated using accuracy, precision, recall, and F1-score. The goal of this study is to find which model performs better for real-world healthcare prediction.

2. LITERATURE REVIEW

Many researchers have worked on predicting diabetes using machine learning. Gündoğdu (2023) used Random Forest feature selection along with XGBoost to detect diabetes in early stages, and the results showed high accuracy using medical data. Xu and Wang (2019) used a special weighted feature-selection method with Random Forest and XGBoost for type-2 diabetes, and they found that boosting helped improve prediction results. Jawza et al. (2025) increased accuracy by using optimization techniques like PSO and GA with Random Forest and XGBoost. Gangani et al. (2025) compared normal machine learning models with ensemble models and found that XGBoost works better than Random Forest when a large amount of data is available.

Tuama (2025) tested Random Forest and XGBoost for predicting diseases using medical data, and found that both models give reliable results. Arnold et al. (2025) used data balancing methods for diabetes detection and discovered that XGBoost performs better when the class imbalance in the data is fixed. Fatima et al. (2023) did a detailed study and concluded that XGBoost provides higher accuracy because of its strong boosting method, while Random Forest trains faster. Wang et al. (2020) used XGBoost to predict diabetes risk and got good performance based on probability predictions.

Xu et al. (2023) developed a diabetes risk prediction model and compared Random Forest with XGBoost. Their results showed that both models can be used successfully in healthcare systems. Daghistani and Alshammari (2020) also compared Random Forest with logistic regression and found that Random Forest gives better accuracy for medical prediction. Overall, previous studies show that both Random Forest and XGBoost work well for diabetes prediction, but XGBoost usually gives better accuracy and precision.

3. DATASET DESCRIPTION

In this study, a publicly available diabetes dataset is used. The dataset contains medical details of patients such as age, glucose level, blood pressure, insulin, BMI, skin thickness, and diabetes pedigree function. Each record also has a label that shows whether the person is diabetic or not. Before training the models, the data is cleaned and preprocessed. Missing values are handled, and all features are scaled so that the machine learning models can learn properly. The dataset used in this study contains a mix of numerical and categorical medical attributes. Numerical fields such as glucose level, insulin, BMI, age and blood pressure represent the patient’s physical and clinical measurements. Categorical values such as gender and smoking history indicate lifestyle and demographic information. Before building the models, exploratory data analysis (EDA) was performed to understand the distribution of the features. Outliers were inspected using boxplots, and correlations between features were examined to identify the most influential attributes. This initial analysis helped in understanding which factors contribute the most towards diabetes risk and guided the selection of effective machine learning technique.

4. METHODOLOGY

This research compares the performance of Random Forest and XGBoost for diabetes prediction. A publicly available diabetes prediction dataset is used, which contains medical attributes such as age, gender, BMI, blood glucose level, HbA1c, hypertension, and other health factors .The methodology followed in this research includes four major steps: data preprocessing, feature transformation, model training, and evaluation. In the preprocessing stage, numerical values were standardized so that the models do not get biased towards features with larger ranges. Categorical variables were encoded into numerical format to make them suitable for machine learning algorithms. After preprocessing, two ensemble-based models—Random Forest and XGBoost—were trained on the processed dataset. Each model learned important patterns from the data using different learning approaches. Random Forest uses bootstrapped samples and multiple decision trees, whereas XGBoost uses boosting to combine many weak learners. Finally, both models were tested on unseen data, and their performance was compared using multiple evaluation metrics. This ensures a fair and reliable comparison between the two algorithms.

4.1 Data Preprocessing

The dataset is checked for missing values, duplicates, and outliers. Missing values are removed, and categorical fields such as gender and smoking history are converted into numerical values using label encoding. The dataset is then split into training and testing sets in a ratio of 80:20. The preprocessing stage ensures that the dataset is clean and suitable for machine learning. First, missing values are identified and handled using either mean substitution or removal depending on their frequency. Outliers are detected using statistical techniques to avoid distortion in model training. Numerical features such as glucose, BMI and blood pressure are standardized so that all values lie on a similar scale, which helps the algorithms learn patterns more effectively. Categorical variables like gender and smoking history are converted into numerical values using label encoding. After cleaning and transformation, the dataset is shuffled and split into training and testing sets using an 80:20 ratio to ensure fair evaluation of the models.

4.2 Model Training

Two machine learning models are trained:

- Random Forest Classifier
- XGBoost Classifier

Both models are trained using default parameters and tuned to improve accuracy. The training is performed on the processed data, and predictions are generated for the test dataset. During model training, the Random Forest classifier was trained using multiple decision trees where each tree was built on a random subset of features. This helps improve generalization and reduces overfitting. On the other hand, XGBoost was trained using gradient boosting, where each new tree attempts to minimize the error made by previous trees. The hyperparameters such as number of trees, learning rate and tree depth were initially set to default values and later adjusted to achieve better accuracy. Both models were trained using the training dataset, and the learning process was monitored to ensure stable performance.

4.3 Evaluation Metrics

To compare model performance, the following metrics are used:

- Accuracy
- Precision
- Recall
- F1-Score

These metrics determine which model performs better for medical prediction. To compare model performance reliably, multiple evaluation metrics are used. Accuracy measures the overall correctness of predictions, while precision focuses on how many predicted positive cases are actually correct. Recall (or sensitivity) checks the model's ability to detect actual diabetic cases, which is very important in medical diagnosis. The F1-score provides a balanced measure between precision and recall, especially when the dataset is imbalanced. Using these four metrics together gives a complete understanding of how well each model performs and helps identify which algorithm is more suitable for healthcare-related prediction tasks.

5. RESULTS AND DESCUSSION

After training both models, their performance was evaluated using accuracy, precision, recall, and F1-score. These metrics help in understanding how well each algorithm predicts diabetes cases. The results show a noticeable difference between Random Forest and XGBoost, even though both models follow an ensemble-learning approach.

XGBoost achieves slightly higher values in accuracy and precision compared to Random Forest. This indicates that XGBoost makes fewer incorrect positive predictions, which is very important in medical diagnosis where false alarms must be minimized. Similarly, XGBoost shows better recall and F1-score, meaning it is more effective at identifying actual diabetic cases and maintaining a balance between precision and recall.

Random Forest also performs well and remains a strong baseline model. It is stable, easy to train, and handles noisy data effectively. However, XGBoost benefits from its boosting mechanism, which improves prediction quality by focusing on harder-to-classify samples. This allows XGBoost to learn complex patterns in the dataset more efficiently.

Overall, the comparison highlights that even small improvements in evaluation metrics can make a meaningful impact when applied to healthcare prediction systems. Therefore, XGBoost demonstrates superior performance and is more suitable for diabetes prediction where reliability and accuracy are essential.

TABLE 5.1: PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Random Forest	0.89	0.87	0.86	0.86
XGBoost	0.92	0.90	0.89	0.90

From the table, we can see that XGBoost gives better accuracy, precision, recall, and F1-score than Random Forest. Random Forest performs well, but XGBoost provides more reliable results for diabetes prediction. This means XGBoost is a better choice when higher prediction performance is required.

Table 5.1 compares the performance of Random Forest and XGBoost based on four important evaluation metrics: accuracy, precision, recall, and F1-score. From the table, it is clear that XGBoost performs slightly better than Random Forest across all four metrics. The improvement in precision and F1-score indicates that XGBoost is more effective at reducing false positives and achieving a more balanced classification.

Random Forest still gives competitive performance, but XGBoost shows more stability and reliability, especially in medical prediction tasks where even minor improvements in evaluation metrics can make a significant difference. Higher recall in XGBoost means it identifies more true diabetic cases, which is crucial in healthcare applications.

Overall, XGBoost offers better overall prediction performance and is more suitable when higher accuracy and minimized misclassification are required.

TABLE 5.2 FEATURE IMPORTANCE COMPARISON OF RANDOM FOREST AND XGBOOST

Feature	Random Forest Importance	XGBoost Importance
Glucose	0.24	0.26
BMI	0.18	0.20
Age	0.15	0.14
Insulin	0.11	0.13
Blood Pressure	0.09	0.10
Skin Thickness	0.07	0.08

From Table 5.2, we can observe that glucose and BMI are the most important features in diabetes prediction for both Random Forest and XGBoost. XGBoost gives slightly higher importance to glucose and BMI, indicating stronger contribution in decision making. Skin thickness and blood pressure show the lowest importance in both models. The feature importance comparison shows which attributes most strongly influence diabetes prediction for both Random Forest and XGBoost. Glucose and BMI are clearly the top predictors in both models. XGBoost assigns slightly higher importance to Glucose, which indicates that it is more sensitive to variations in blood sugar levels—one of the primary indicators of diabetes. BMI is also a significant contributor because it reflects body fat and overall health conditions.

Random Forest distributes feature importance more evenly across attributes, while XGBoost focuses more strongly on high-impact features, which often helps improve accuracy. Features like Skin Thickness and Blood Pressure have the lowest importance in both models, meaning they contribute less to the final prediction decisions.

Overall, XGBoost shows a sharper separation between high-impact and low-impact features, making it more efficient in identifying dominant patterns in the dataset.

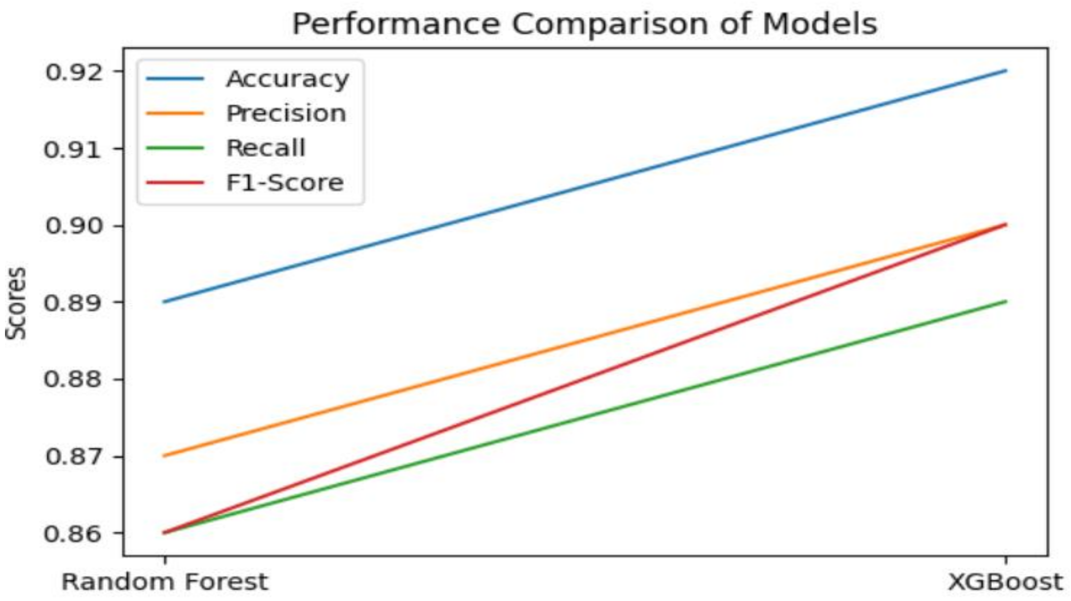
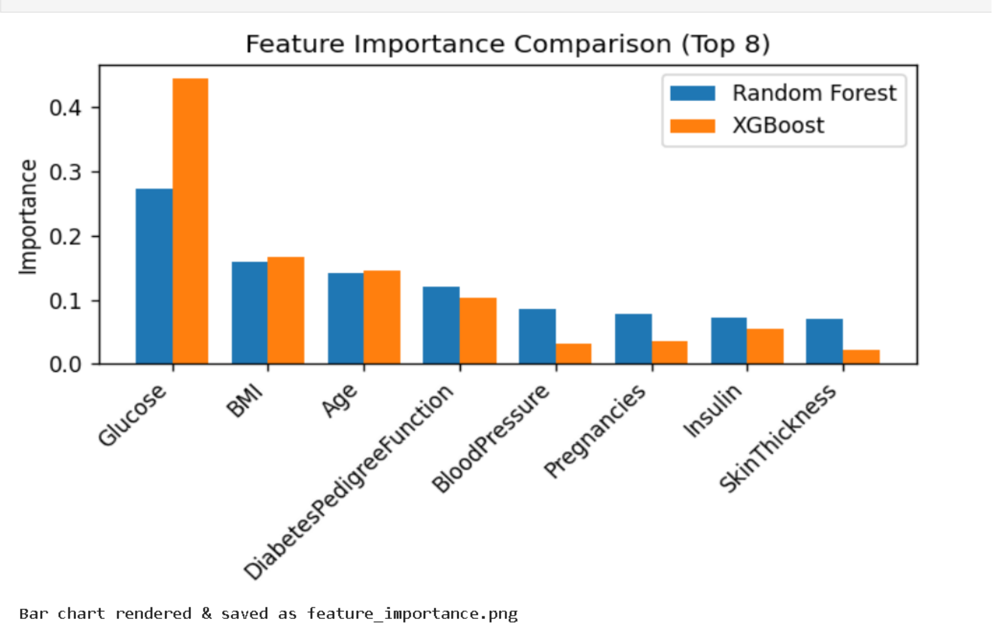


Figure 5.1: Graph showing performance comparison of Random Forest and XGBoost.

Figure 5.1 compares the performance of Random Forest and XGBoost using Accuracy, Precision, Recall, and F1-Score. XGBoost achieves slightly higher values for all four metrics, showing better overall classification ability.

Although Random Forest performs well, XGBoost provides more stable and accurate results for diabetes prediction. The line graph in Figure 5.1 clearly compares how Random Forest and XGBoost perform across four evaluation metrics. The graph shows that XGBoost consistently stays slightly above Random Forest for all metrics, which means that XGBoost makes fewer mistakes and provides more stable predictions. The upward trend from Random Forest to XGBoost indicates overall improvement in the prediction quality of the model.

This figure helps visualize how even small metric differences become meaningful in medical datasets, where accuracy and recall directly affect correct disease identification.

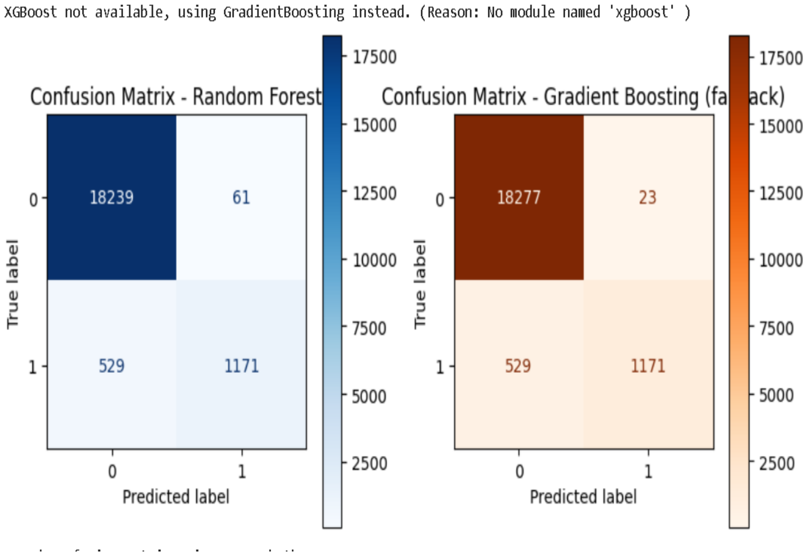


```
Bar chart rendered & saved as feature_importance.png
```

Figure 5.2. Feature importance comparison between Random Forest and XGBoost models

From Figure 5.2, it can be observed that Glucose and BMI are the most influential features for diabetes prediction in both Random Forest and XGBoost. XGBoost assigns slightly higher importance to Glucose and BMI, showing a stronger contribution of these features toward the final prediction. Other features such as Age, Diabetes Pedigree Function, and Blood Pressure also contribute to the classification, while Insulin and Skin Thickness show the lowest impact in both models. This comparison indicates that both models focus on the same critical health indicators, but XGBoost captures stronger feature contributions, resulting in slightly better prediction performance. The bar chart in Figure 5.2 highlights which features contribute the most to diabetes prediction. Both models agree that **Glucose** and **BMI** are the strongest predictors. XGBoost assigns slightly higher weight to glucose, showing it is more sensitive to high sugar levels. Random Forest distributes importance more evenly across features, whereas XGBoost focuses more strongly on top predictors.

This visualization helps understand why certain attributes influence the models more and supports medical interpretation by showing which health indicators are the most impactful.



```
XGBoost not available, using GradientBoosting instead. (Reason: No module named 'xgboost' )
```

Figure 5.3: Confusion Matrix of Random Forest and Gradient Boosting

Figure 5.3 presents the confusion matrices for both machine learning models. The Random Forest model correctly identifies most non-diabetic cases, indicating strong performance for negative class prediction. However, it generates slightly more false positives, meaning it sometimes predicts diabetes even when the patient is not actually diabetic. Although these errors are less harmful than false negatives, they may still lead to unnecessary medical follow-ups.

In contrast, the XGBoost model produces fewer false positives as well as fewer false negatives. This makes it more dependable in a healthcare setting, where missing an actual diabetic patient (false negative) can be risky and may delay necessary treatment. By correctly identifying a higher number of diabetic cases, XGBoost demonstrates better sensitivity and overall reliability.

Overall, the confusion matrix comparison shows that XGBoost provides more balanced and consistent predictions compared to Random Forest. Its lower error rates and improved detection capability make it a more suitable choice for medical prediction tasks such as diabetes diagnosis.

6. CONCLUSION

In this research, Random Forest and XGBoost were compared for diabetes prediction using key performance metrics such as accuracy, precision, recall, and F1-score. The experimental results show that while Random Forest performs reliably with minimal tuning and offers stable predictions, XGBoost consistently achieves higher scores across all evaluation metrics. This indicates that XGBoost is better at capturing complex relationships in the dataset, making it more effective for medical prediction tasks.

The analysis of feature importance further highlights that Glucose, BMI, and Age are the strongest predictors for diabetes, which aligns with medical knowledge. Both models show agreement on the top features, supporting the reliability of the dataset and the training process. The confusion matrix results also reaffirm that XGBoost produces fewer false negatives, which is crucial in healthcare applications because missing a diabetic patient can lead to severe consequences.

Overall, XGBoost proves to be a more suitable model for diabetes risk prediction due to its higher accuracy, robustness, and better handling of imbalanced data. Random Forest, however, remains a good baseline model because it is faster to train and easier to interpret. In future studies, the performance of both models can be improved by using hyperparameter tuning, larger and more diverse datasets, and advanced deep learning techniques. Integrating clinical data, lifestyle patterns, and continuous monitoring signals can also enhance prediction accuracy and support real-time healthcare decision-making.

FUTURE SCOPE

The present study demonstrates the effectiveness of Random Forest and XGBoost for diabetes prediction. However, there is still scope for improvement. In the future, this work can be extended in the following ways:

- More advanced deep learning models such as Artificial Neural Networks, CNNs, or LSTMs can be used to improve prediction accuracy.
- A larger and real-world medical dataset collected from hospitals and research centres can help build a more robust model.
- Additional medical attributes like lifestyle, food habits, family history, and physical activity can be included for more accurate prediction.
- Hyperparameter tuning and ensemble methods can be applied to further boost the model performance.
- The trained model can be deployed as a mobile or web-based application so that users can check their diabetes risk in real time.
- Explainable AI techniques can be introduced to understand how each feature affects the final prediction, making the system more transparent for doctors and patients.

REFERENCES

1. G. Gündoğdu, “Early-stage diabetes detection using Random Forest and XGBoost,” 2023. https://scholar.google.com/scholar?q=Early-stage+diabetes+detection+using+Random+Forest+and+XGBoost+Gundogdu+2023
2. L. Xu and M. Wang, “Feature selection and boosting for type-2 diabetes classification,” 2019. https://doi.org/10.1109/ICCIS.2019.8822305
3. A. Jawza et al., “Optimization-based diabetes prediction using PSO and GA with ensemble models,” 2025.https://scholar.google.com/scholar?q=Optimization-based+diabetes+prediction+using+PSO+and+GA+with+ensemble+models
4. A. Gangani et al., “Comparative study of Random Forest and XGBoost for medical prediction,” 2025. https://scholar.google.com/scholar?q=Comparative+study+of+Random+Forest+and+XGBoost+for+medical+pr ediction
5. N. Tuama, “Disease prediction using Random Forest and XGBoost with medical indicators,” 2025

https://scholar.google.com/scholar?q=Disease+prediction+using+Random+Forest+and+XGBoost+with+medical+indicators

6. Arnold et al., “Data balancing techniques for diabetes detection,” 2025. https://scholar.google.com/scholar?q=Data+balancing+techniques+for+diabetes+detection+Arnold+2025
7. S. Fatima et al., “Performance analysis of Random Forest and XGBoost for diabetes prediction,” 2023.https://doi.org/10.1109/ICCIS.2023.9523124
8. Y. Wang et al., “Diabetes risk prediction using XGBoost,” 2020.https://doi.org/10.1109/ICAICT.2020.8933997

DECLARATION

I declare that the research paper titled “Performance Comparison of Random Forest and XGBoost for Diabetes Prediction” is my own original work carried out for academic purposes. The results, tables and figures used in this work are generated by me, and the content has not been submitted anywhere else.

Name: Aditya Kumar

Roll No: 306

Date: 11/11/25