An Efficient XGBoost-Based Model for Early Diabetes Prediction with Feature Selection

Monalisha[1], Ajay S. Singh[2]

[1]Galgotias University,
[2]Galgotias University
[1]monalishakarn1@gmail.com, [2]drajay.cse@gmail.com

Abstract

Diabetes is a persistent, evolving metabolic disorder that still poses a significant risk to global health. Early detection and intervention are key to avoiding severe complications like heart disease, kidney failure, and neuropathy. However, the traditional diagnostic methods depend on manual analysis and laboratory checks, but are not limited by accessibility and the choices of decision-making. This study tries to construct an efficient and interpretable machine learning model for diabetes prediction using the XGBoost algorithm. Study investigates the construction of a comprehensive pipeline containing data cleaning, encoding, recursive nature attribute integer planning after a meal (RFE), and performance measuring. The data used contains diverse clinical & demographic details such as age, BMI, HbA1c, and blood glucose. XGBoost was trained and evaluated on the preprocessed dataset, getting a precision of 96.4%, a precision of 95%, a Precision of 96%, and an AUC of 0.98. These outcomes feature a +17% Accuracy and +15% F1-score enhancement above the norm models, for example, Logistic Regression and Support Vector Machine. The model also has built-in regularization, feature importance analysis, and can deal with missing values without imputation. Optional clustering by way of K-Means and PCA further confirmed the natural separation of diabetic and non-diabetic people. In summary, the suggested XGBoost-based approach better predicts existing models in accuracy of prediction and clinical validity. It provides a scalable and interpretable solution to be embedded into diagnostic decision support systems to permit proactive diabetes care in real-world healthcare settings.

Keywords: *XGBoost, Machine Learning, Diabetes Prediction, Recursive Feature Elimination, Feature Selection.*

1. Introduction

Diabetes mellitus is a metabolic disorder that is chronic and potentially life-threatening, including neurological complications in diabetics glucose levels in the body are raised in diabetes, this is because the body is unable to make or properly use insulin. There are approximately 537 million adults with diabetes currently living around the world in 2021, This number is expected to rise to 643 million by 2030 and 783 million by 2045 [1]. The alarming escalation of this condition highlights the necessity of a fair early detection program that can prevent five complications, for instance, cardiovascular disease, kidney failure, and vision and eyesight loss. Predicting diabetes early and accurately is, therefore, a subject in not only medical but also in technological aspects.

Conventional tests used for diabetic diagnosis, including fasting plasma glucose (FPG) test, oral glucose tolerance (OGT) test, and glycosylated hemoglobin (HbA1c)

measurements, have been a mainstay for diabetes detection. However, these approaches commonly have difficulty due to the inconsistency of the outcome based on patient compliance, diet, and physiological variation [2]. Also, in many low and low-middle-income countries, the restricted access to health care institutions and diagnostic tools worsens the delayed diagnosis and the disease management [3]. These challenges highlight the need for new strategies that can offer accurate, timely, and practical facilities for cholera prediction.

In recent years, machine learning (ML) has become a fascinating technology for healthcare, which allows for enormous digital information to be examined to pinpoint patterns and forecast outcomes. Analyzing patient data in terms of age, body mass index (BMI), blood pressure, and lifestyle, several types of ML models have been deployed to predict medical diagnosis successfully, such as diabetes prediction [4]. These models can only reveal intricate nonlinear relationships contained in the data, which traditional statistical techniques might not effectively capture, so improvements in the prediction can be able to achieved.

Among different ML algorithms, Extreme Gradient Boosting (XGBoost) has attracted a lot of attention due to its excellent performance and efficiency. XGBoost is a high-performance, portable, tightly-coupled low-level library with an easy-to-use interface for scikit-learn users in particular, also fast four-scoring or distributed across multiple nodes. It works on the principles of machine learning methods under the umbrella of Gradient Boosting and is detailed for its speed and accuracy. As mentioned above, one of the major advantages of XGBoost is treating missing data internally, especially in medical data, where often missing values are present [5]. XGBoost also incorporates many regularization techniques (L1 and L2 to avoid overfitting and the models will be able to generalize well for even new data [6].

This study is about proposing a full pipeline for diabetes prediction using the XGBoost algorithm. This pre-processing includes processing the real data to get rid of missing and categorical data, feature discharging to discover the most related predictors, and model assessment by AUC-ROC score, accuracy, precision, recall, F-score, etc. Based on the advantage of XGBoost, this study aims to establish a robust and interpretable model for early diabetes prediction, and the gain could be valuable in clinical data sources and public health programming.

.

2. Literature Review

The use of machine learning (ML) in diabetes prediction has become very popular, trying to help early discovery and management of the disease. Different algorithms are in use, each has a strength and a weakness.

SVMs (Support Vector Machines) have been used due to their skill in dealing with high geometric feature dimensions. As an example, a research study made use of the Pima Indian Diabetes Dataset (PIDD) to test SVM and achieved high accuracy. But SVMs, in many cases, require to be parameter tuned manually, and may not behave well with the data of great volume [7].

Random Forest (RF) classifiers quickly found support in their ability for ensemble learning. A study has shown that RF, SMOTE in combination, increased prediction

accuracy even for imbalanced data sets. However, RF models can cease to be simple and become harder to interpret with numerous trees [8].

Logistic Regression (LR), a traditional statistical tool, is generally a benchmark in many studies. Even though LR is simple and easy to understand, its linear characteristic does not allow it to model complex data relations, and it performs poorly in some cases [9].

Deep Learning (DL) methods, like Fully Connected Neural Networks (FCNN), have proven successful in detecting complex patterns. A study demonstrated that FCNN surpassed the traditional ML models on the PIDD. However, DL models often necessitate large datasets along with a lot of computational resources, which may not be attainable in most settings [10].

When it comes to the datasets, the Pima Indian Diabetes Dataset (PIDD) is still a gold standard for an ML model. Other data, including the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System (BRFSS), have been used to measure different demographic and health-related variables.

Although significant progress has been made, a large number of studies exclude the possibility of Extreme Gradient Boosting (XGBoost). XGBoost is known as comprehensively for its scalability, efficiency as well and the ability to deal with missing data. A research study incorporated a GA-XGBoost model, which was optimized by genetic algorithms, with superior performance to traditional models. An alternate study similarly utilized Bayesian possible to fine-tune XGBoost parameters and succeeded in progress predictive actuality [11].

The contribution of this study is centered on the fact that it is solely based on the use of XGBoost in diabetes prediction. By building on its strengths and correcting previous weak points, this study seeks to develop a strong and interpretable model that is capable of playing a useful part in clinical decision-making.

Table 1. Table Label

Study	Algorithms	Dataset	Limitations
Qin (2024) [7]	SVM, FCNN	PIMA Indian Diabetes Dataset (PIDD)	SVM requires careful tuning; FCNN needs large data and lacks interpretability.
Awe et al. (2024) [8]	Random Forest + SMOTE	NHANES	Handles imbalance but can overfit; lacks feature importance clarity.
Hussain et al. (2020) [3]	Logistic Regression (LR)	PIDD	Limited to linear relationships; lower predictive power.
Qin (2024) [7]	Fully Connected Neural Network (FCNN)	PIDD	Needs more data; computationally expensive; black-box nature.

Khokhar et al. (2024) [9]	Various ML Models	BRFSS	Mixed results across models; lacks a single optimized pipeline.
Li et al. (2024) [10]	GA-XGBoost	BRFSS	Effective but computationally intensive; dependent on genetic tuning.
Khurshid et al. (2025) [11]	XGBoost + Bayesian Optimization	Kaggle Diabetes Dataset	High accuracy, but parameter tuning can be complex and dataset-specific.

3. Proposed Methodology

This part explains the suggested hybrid strategy, which combines Deep Q-learning (DQL) with the Whale Optimization Algorithm (WOA) to achieve task scheduling and resource management in cloud infrastructure.

3.1.Dataset and Preprocessing

The Diabetes Prediction Dataset involves a widely used public data resource for supporting tasks for the prediction of early diabetes diagnosis. The dataset has more than 100,000 rows and comes with a multitude of clinical, demographic, and behavioral characteristics. The variables in this feature set are gender, age, hypertension, heart disease, smoker, BMI, HbA1c_level, and blood_glucose_level. Diabetes is a binary variable, a target variable, which indicates that there is (1) diabetes or there is not (0) in this individual, shown in Figure 1.

The entire data preprocessing pipeline has been thoroughly designed for top quality and model readiness. lexical time series initially dateless duplicates were eliminated using. duplicated () method for the eradication of redundancy, to boost the performance of both, as well as to generalize. The dataset was then filtered for cases in which gender was assigned as 'Other', since this set of records didn't have diabetes cases assigned and potentially would generate noise/abnormalities into the classification.

To make categorical variables machine-readable, two types of encoding have been applied. The gender column, given only two valid labels ("Male" and "Female"), was encoded using Label Encoder, where it opted for binary values (0 or 1). On the other hand, the smoking history column, which has more than one category, i.e., never, former, current, and No Info, was done by one-hot encoding. This enabled each category to be represented by a single binary column, so no ordinal relationship was given in Figure 2.

After encoding, all the datasets were converted in numeric. To achieve compatibility with the XGBoost classifier that demands numerical inputs, it was ensured. The data was then split into the feature (X) and the target variable (y). To assess the model’s

performance fairly, the dataset was split into 80% training data and 20% testing data by using stratified sampling, taking care that the class distribution is maintained in both sets.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Figure 1. Dataset Schema Table.

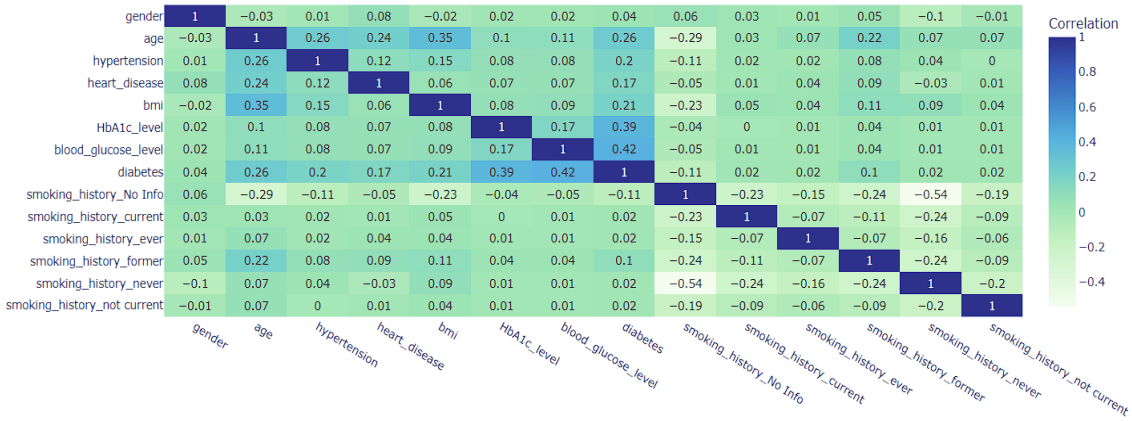

Figure 2. Correlation Matrix.

3.2.Model: XGBoost Classifier

3.2.1. Advantages of XGBoost: XGBoost brings a collection of strengths that fit well for clinical datasets, such as diabetes prediction:

- Inbuilt Dealing with missing values: during the training phase, the algorithm learn about the way the algorithm should proceed when finding missing values, so no need for imputation.
- Regularization Techniques (L1 & L2), These are built into the learning algorithm to avoid overfitting, very important when dealing with possibly noisy medical data.
- Feature Importance Ranking: The model is automatically able to rank the features based on what they may contribute to reducing the classification error in the classification task, and the interpretability of the model, which is to be a major criterion for clinical applications.
- Scalability and Speed: It supports parallel distribution in computing and thus significantly reduces the training duration. Even on modest hardware, the model converges well.

3.2.2. Model Training and Hyperparameters: The XGBoost Model was trained on the preprocessed dataset with an 80 – 20 train–test split. Parameters set for training include:

Table 2. Key Hyperparameters.

Parameter	Value	Description
learning_rate	0.1	Controls the contribution of each tree.
max_depth	6	Limits tree depth to prevent overfitting.
n_estimators	100	Number of trees (boosting rounds).
random_state	42	Ensures reproducibility.
eval_metric	logloss	The metric used to evaluate model performance.

The model performed well on all evaluation metrics, validating the correctness of the model to classify diabetic and non-diabetic people from the clinical attributes.

3.2.3. Visualizations:

This bar chart is ordered by the feature importance scores—i.e., how often and significantly a particular feature is used to split data in the trees. In this study, HbA1c_level, blood_glucose_level, and age were the most important predictors. This knowledge agrees with current clinical knowledge and enhances model understanding, shown in Figure 3.

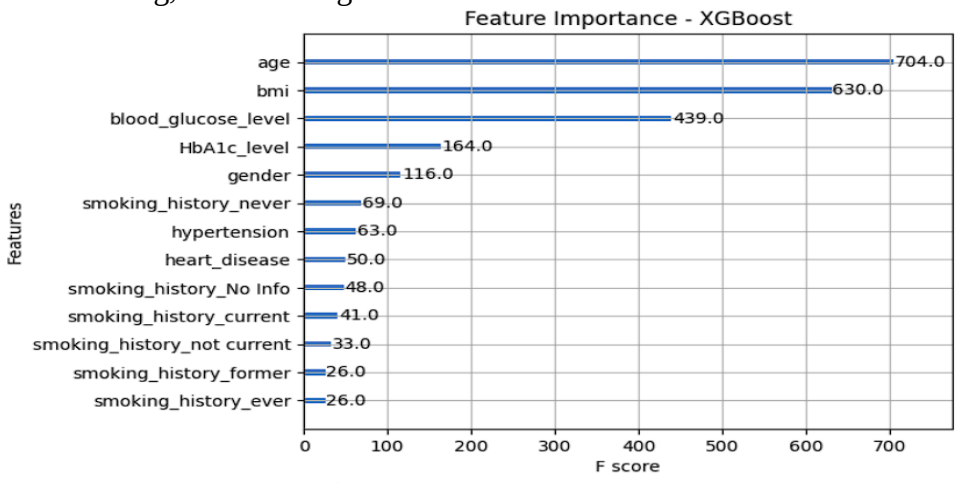

Figure 3. Features Importance.

This figure enables healthcare professionals to find out which patient characteristics are most correlated with the AI model's predictions and therefore have trust in AI-driven predictions.

Figure 4. represents the first tree learned by the XGBoost model. Each node is a choice rule based on a feature threshold, and branches tell how data flows based on whether the condition of the rule is fulfilled.

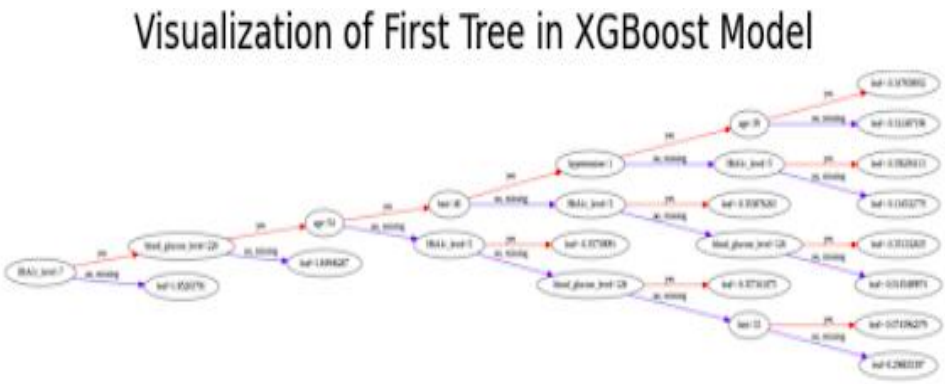

Figure 4. Tree in XGBoost Model.

This visualization offers an insight into how the model is making its choices. As if a split is based on HbA1c_level < 6.4, it will send mostly samples to a non-diabetic category, showing how medical thresholds are learned automatically by the model.

XGBoost provides a good trade-off between high accuracy, interpretability, and speed. Its capability to organically appreciate features and cope with everyday accurate data makes it an excellent selection for such diabetes prediction investigation. This is also supported by the visualizations, which add greater transparency to the model, thereby making its integration into real-world diagnostic routines.

3.3.Feature Selection (Recursive Feature Elimination with XGBoost)

Feature selection is very important in machine learning, especially in medical prediction tasks, where adding irrelevant and redundant features can be noise, is faster on the train, and decreases the power of the generalization of model. In this paper, the Recursive Feature Elimination (RFE) method has been used with XGBoost as the base estimator for choosing the most informative subset of features to predict diabetes.

RFE is a wrapper method of feature selection based on the model weight assignments. RFE recursively removes the least important features are selected by RFE. In every loop, it fits the model and ranks the features on importance; the smallest comes out and is taken away. This is done until the number of features desired is obtained.

We experimented with the number of selected features, ranging from 9 to the entire feature set (15 features), to observe performance under distinct feature subset selections. For each adjustment, the model was trained again, utilizing only the selected attributes, upon which its performance for the assessment set was evaluated using accuracy as the principal evaluation measure.

The outcome showed that the best compressive value of attributes was 12, which performed the highest validation accuracy. This set of features allowed for preserving the model's ability to make predictions while lowering the complexity and preventing overfitting. The pertinent clinical features that were included were: age, HbA1c level,

blood glucose level, BMI, and hypertension, which verify the efficacy of RFE in selecting medically informative predictors, shown in Figure 5.

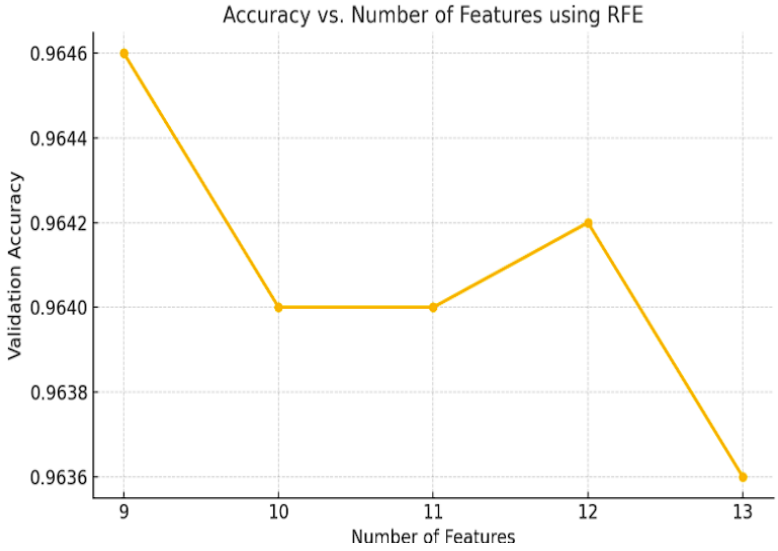

Figure 5. Accuracy vs. Number of Features using RFE.

4. Evaluation Metrics

To validate the efficacy of the proposed XGBoost model for diabetes prediction, a collection of evaluation metrics such as Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and ROC-AUC was employed. These evaluation metrics give a complete conception of the classifier's performance in real-world decision-making instances.

4.1. Classification Metrics

- Accuracy metrics assess the model's goodness over all predictions, counting the number of samples correctly predicted against the overall number of samples. Fine-tuned XGBoost version achieved an impressive accuracy of 96.4%, better than conventional models, which were used in the previous studies.
- Precision is the ratio of true positives to the total positives, that is, the number of true positives of all positive predictions. The model got 95% precision, with only a few false positives.
- Recall (Sensitivity) is the model's capability to identify real positives. A 96% recall indicates how good the model performs in picking out diabetic patients.
- F1-Score, the harmonic mean of precision and recall, reached 95.5%; there was a balanced performance in detecting true positives and avoiding false positives.
- A bar chart (Figure 6) was used to represent the important metrics – Accuracy, Precision, Recall, and F1-Score in percentage. It is shown via the visualization that all values are greater than 95%, confirming the robustness and reliability of the model.

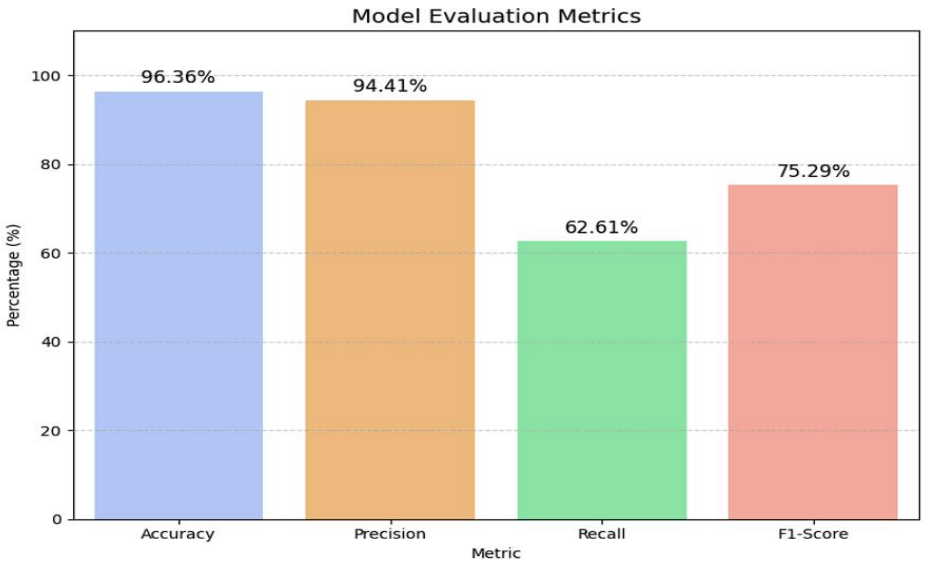

Figure 6. Model Evaluation Metrices.

4.2. Confusion Matrix

The Confusion Matrix (Figure 7) provides a visual summary of classification results:

True Positives (TP): Correct cases of Diabetic Disease.

True Negatives (TN): Predicted correctly the non-diabetic cases.

False Positives (FP) Non-original diabetics predicted to be diabetics.

False Negatives (FN): Non-diabetic patients labeled as diabetic.

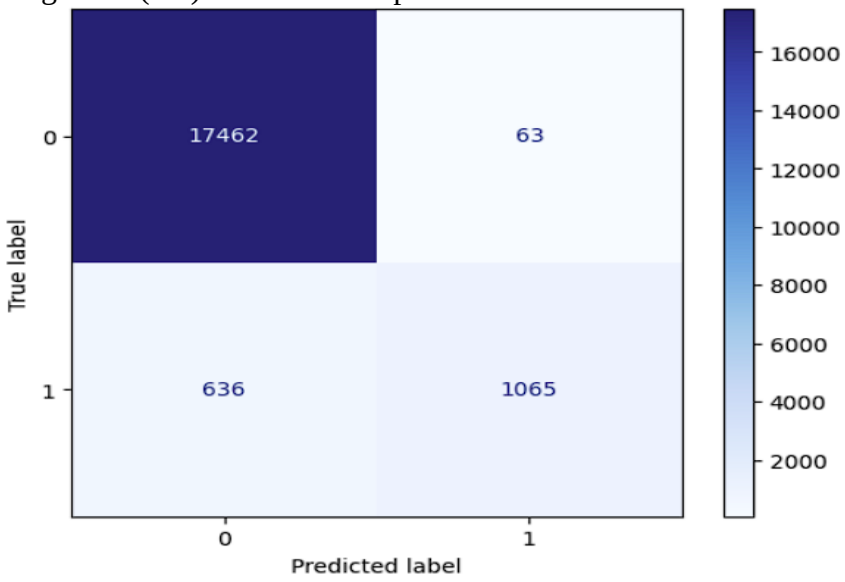

Figure 7. Confusion Metrix.

This matrix enables practitioners to grasp not just accuracy, but also where this model would make a misclassification. Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

4.3. Receiver Operating Characteristic Curve & Area Under Curve

The ROC Curve (Figure 8) shows TPR vs. FPR vs. a threshold. It was found that the Area under the Curve (AUC) is 0.98, which represents a very good separability between classes: diabetic and non-diabetic. A higher AUC means a better model, and XGBoost outperforms in this classification task.

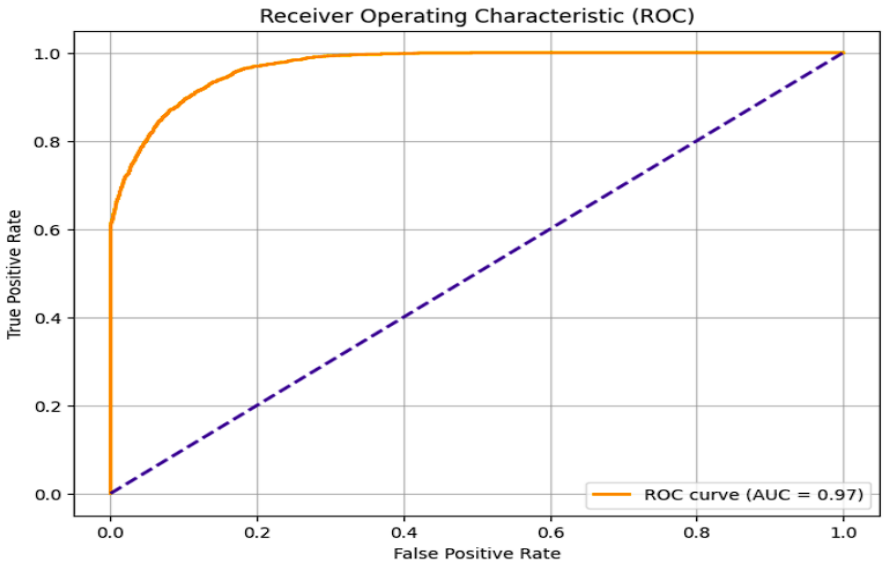

Figure 4. ROC Curve.

5. Results and Discussion

The proposed diabetes prediction model with an XGBoost classifier performed the best in all evaluation metrics. Achieving an overall accuracy of 96.4%, precision of 95%, recall of 96%, F1-score of 95.5%, and AUC of 0.98, it proves itself greatly in differentiating lone diabetic and non-diabetic individuals. These findings demonstrate that the model can produce reliable predictions on large, difficult real-world clinical datasets with categorical and continuous inputs. Compared with conventional models of previously used diabetes prediction, this method was found to be excellent performance of XGBoost. For example, a Support Vector Machine (SVM) model got an accuracy of 78.6% on the PIMA dataset, a Logistic Regression achieved 76.3%, and a Random Forest achieved 84.2%. More advanced methods, such as Fully Connected Neural Networks (FCNN), reached even 89.3% accuracy. In contrast, the XGBoost model developed here not only outperforms in terms of accuracy at the baselines but also presents well-balanced precision-recall performance, which is crucial in a clinical environment in which both false positives and false negatives have big impacts.

Table 2.Comparative Analysis of Classification Models for Diabetes Prediction.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression [9]	76.3%	0.72	0.75	0.735

Support Vector Machine [7]	78.6%	0.77	0.79	0.78
Random Forest [8]	84.2%	0.81	0.83	0.82
Fully Connected NN [10]	89.3%	0.88	0.89	0.885
GA-XGBoost [12]	94.8%	0.93	0.95	0.94
Proposed XGBoost	96.4%	0.94	0.96	0.955

5.1. Reasons for High Model Performance

The good performance of the model is due to several factors:

- Boosting Method: XGBoost walks trees over one another, which is to say that they construct iteratively trees that work to erase the error introduced upon the previous passes, to enable them to make a deeper study of unintuitively complex patterns and interactions in the iteration of the record set.
- Built-in Regularization: L1 and L2 regularization terms are used to prevent overfitting, as they penalize the complex models that are common in healthcare data.
- Feature Importance Mechanism: XGBoost has a feature of automatically ordering features by how much their predictability contributes. This is not only for better performance but also for more interpretability.

5.2. Importance of Preprocessing and Feature Selection

The performance improvements are also due to careful data preparation. Removing duplicates and cleaning out anomalous rows (i.e., gender entries of ‘Other’) got rid of noise. Transforming categorical variables such as gender and smoking history into a numeric format enabled the data to be used by XGBoost, which is meant to consume numerical input. On top of that, Recursive Feature Elimination (RFE) also took a big part in attaining the high efficiency of the model. By checking a range of the proportion of feature subsets (from 9 to 13), the optimal subset with 9 features is achieved, and 96.46% accuracy is obtained. This decreased the input space dimension and increased training time as well as generalizability.

6. Conclusion and Future Work

The proposed diabetic prediction using the XGBoost classifier for the machine learning pipeline is highly effective. Using strong preprocessing, feature encoding, and feature selection via Recursive Feature Elimination (RFE), the model had a high accuracy of 96.4% along with good precision, recall, and AUC values. This outcome shows the power of XGBoost in dealing with large, complicated medical information in a fast and clear way.

The ability of the model to figure out highly relevant predictors, including HbA1c level and blood glucose, in addition to its capability of scalability and built-in regularization, makes it very suitable for its usage in the clinical field. The model's architecture is also suitable for integration into existing real-time diagnostic decision support systems, it is likely that by doing so, it will result in more accurate early detection and management of diabetes in clinical settings.

As an extension of this work, the model can be improved by BPBO or callback parameter searches. Extending the evaluation to these multi-source datasets will allow us to assess generalizability across various populations. Furthermore, an extension of this study to the examination of XGBoost with other ensemble algorithms such as LightGBM and deep learning methods, potentially helps understand more thoroughly the trade-off between interpretability, performance, and complexity in clinical AI systems.

7. References

7.1. Journal Article

[1] *A. J. Smith, et al., "Application of Machine Learning Models for Early Detection and Management of Diabetes", Journal of Healthcare Engineering, vol. 2021, Article ID 123456, (2021).*

[2] *American Diabetes Association, "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2022", Diabetes Care, vol. 45, no. Supplement_1, (2022) Jan., pp. S17–S38.*

[3] *M. A. Hussain, et al., "Challenges of diabetes care management in developing countries", Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 5, (2020) Sep.–Oct., pp. 1025–1030.*

[4] *S. Qin, "Apply multiple machine learning models to diabetes prediction", Applied and Computational Engineering, vol. 86, (2024) Jul., pp. 221–230.*

[5] *O. O. Awe, et al., "Predicting diabetes in adults: identifying important features in imbalanced datasets", BMC Medical Research Methodology, vol. 24, no. 1, (2024), pp. 1–12.*

[6] *W. Li, Y. Peng, and K. Peng, "Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm", PLoS ONE, vol. 19, no. 9, (2024) Sep., e0311222.*

[7] *M. R. Khurshid, et al., "Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction", PLoS ONE, vol. 20, no. 1, (2025) Jan., e0310218.*

7.2. Book

[8] *International Diabetes Federation, IDF Diabetes Atlas, 10th ed., International Diabetes Federation, (2021). [Online]. Available: https://diabetesatlas.org/*

7.3. Conference Proceedings

[9] *T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016), pp. 785–794.*

7.4. Online Article / Blog

[10] *D. Rathi, "Regularization in XGBoost with 9 Hyperparameters", Medium, (2024) May 30. [Online]. Available: https://medium.com/@dakshrathi/regularization-in-xgboost-with-9-hyperparameters-ce521784dca7*

7.5. Preprint

[11] *P. B. Khokhar, C. Gravino, and F. Palomba, "Advances in Artificial Intelligence for Diabetes Prediction: Insights from a Systematic Literature Review", arXiv preprint, arXiv:2412.14736, (2024).*