INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

AUTONOMOUS DIABETES DIAGNOSIS USING XGBOOST AND DECISION TREE CLASSIFIERS ON ROUTINE CLINICAL DATA

Muzdalifa Bibi[1], Shahan Yamin Siddiqui[*2], Abdullah Dar[3], Shahzada Atif Naveed[4], Muhammad Waleed[5], Muhammad Farrukh Khan[6], Unaiza Rehman[7]

[1]*Department of Computer Science, Minhaj University, Lahore, Pakistan*
[*2,5] *Department of Computer Science, NASTP Institute of Information Technology, Lahore 54810, Pakistan*
[3] *Global Reporting and Analysis Department, Ibex Global Solution, Lahore, Pakistan*
[4]*Department of Computer Science, Rashid Latif Khan University, Lahore, Pakistan*
[6]*Department of Artificial Intelligence, NASTP Institute of Information Technology, Lahore 54810, Pakistan*
[7]*Department of Software Engineering, NASTP Institute of Information Technology, Lahore 54810, Pakistan*

[1]muzdalifabibi3@gmail.com, [*2]drshahan@niit.edu.pk, [3]abdullah.dar@ibex.co, [4]shahzada.atif@rlku.edu.pk, [5]muhammad.waleed@niit.edu.pk, [6]dr.farrukhkhan@niit.edu.pk, [7]unaizarehman@niit.edu.pk

DOI: https://doi.org/10.5281/zenodo.24130979

Keywords
Diabetes Mellitus, Autonomous Diagnosis, Machine Learning, XGBoost, Decision Tree, Class Imbalance, SMOTE.

Article History
Received: 05 August 2026
Accepted: 19 October 2026
Published: 11 November 2026

Copyright @Author
Corresponding Author: *
Shahan Yamin Siddiqui

Abstract

Diabetes mellitus affects an estimated 537 million adults worldwide, however, many cases go undiagnosed until complications arise. This results in a confirmation that depends on the periodic testing of tissues in a laboratory, whose results have to be interpreted manually, leading to a delay in diagnosis in cases of limited availability of specialists. The implementation of a machine-learning autonomous diagnostic system that assesses the information about patients, which can be found in the records of internal organs and laboratory work, has been highlighted in the present article. The machine-learning diagnostic system is organized as a pipeline, which consists of the processes of data-eaned, data allocation, model fitting, and diagnosis output as a class label with a probability score; thereby, the system returns not only a label of a class but also the corresponding probability score as a measure of the system's confidence. The study used 27,551 anonymized patient records from the publicly available Kaggle Diabetes Prediction Dataset, described by age, gender, hypertension, heart disease, smoking history, body mass index, glycated hemoglobin (HbA1c) and blood glucose, with a diabetic prevalence of about 9%. After duplicate removal, categorical encoding and standardization, a Decision Tree and an XGBoost classifier were trained on 70% of the records and tested on the remaining 8,266. XGBoost achieved an accuracy of 96.7%, a precision of 93.5%, a recall of 69.0% and an F1-score of 79.4%, against 94.9%, 71.6%, 72.8% and 72.2% for the Decision Tree, and it reduced false-positive diagnoses from 217 to 36. In a complementary experiment, a random forest trained with SMOTE oversampling and random under sampling raised recall to 83.3%, and HbA1c and blood glucose together accounted for three-quarters of its feature importance. These findings show that tree-based models built on routine indicators can screen for diabetes accurately with few false alarms, and that

rebalancing the minority class is the most effective lever for improving sensitivity before clinical deployment.

INTRODUCTION

Diabetes mellitus occurs as a chronic metabolic disease defined by continuous elevated blood sugar levels. Presently, it ranks among the major causes of sickness and early death on a global scale. According to the International Diabetes Federation, around 537 million adults were diagnosed with diabetes in 2021, with Pakistan recording the highest rates of diabetes in the world (about 33 million cases, with an adult prevalence of 26.7%) [1]. Due to the length of its asymptomatic stage, many individuals do not receive a diagnosis of diabetes until the development of complications affecting the eyes, kidney, cardiovascular system and nervous system has occurred [1], [2].

The current diagnosis relies on laboratory tests like fasting plasma glucose, the oral glucose tolerance test, and glycated hemoglobin (HbA1c) which need to be manually interpreted against clinical cutoff values. Although these tests are reliable, they offer only a preview of the patient during a single visit, requiring multiple laboratory visits which relies on the clinician to make sense of several parameters [2]. Furthermore, differences between patients add complexity in that some may have raised glucose levels but normal HbA1c values while some others might display early insulin resistance with normal glucose levels. In high-volume or low-resource settings, such factors lead to missed or delayed diagnosis.

Machine learning is a solution for automating parts of this process. Data-driven models have become as competent as clinicians in several narrowly defined medical tasks [3] such as identification of diabetic retinopathy in retinal photographs [4] and classification of arrhythmias based on ambulatory electrocardiograms [5]. A review of 102 machine learning diagnosis studies found that support vector machines (SVM), logistic regression, random forests, decision trees and neural networks dominate the field, and that hybrid and ensemble approaches are generally more accurate than single models [2]. For diabetes screening from routine data, where inputs are structured laboratory and demographic attributes rather than images, tree-based models are particularly attractive because they handle mixed feature types and non-linear interactions with little preprocessing [6], [7].

A large body of work has applied these techniques to diabetes prediction. Classical classifiers evaluated on the Pima Indians Diabetes dataset typically report accuracies of about 80–83%, as with AdaBoost [8], random forest [9] and a voting ensemble of five learners [10], while other studies have examined naive Bayes and linear SVM [11] or logistic regression on national survey indicators [12]. Studies on larger or population-specific data, such as the Behavioral Risk Factor Surveillance System [13], a hospital questionnaire cohort from Sylhet [14] and the recent DiaBD dataset from Bangladesh [15], show that both the achievable accuracy and the most suitable model depend strongly on the population and on the attributes available.

Several limitations recur in this literature. Much of it depends on a small benchmark collected from a single population, so the reported models may not transfer to South Asian patients, among whom prevalence is highest [1]. Although accuracy is frequently reported on its own despite the fact that diabetic cases are often in the minority, this could obscure a low level of sensitivity; a review reported that neither class imbalance nor interpretability was given much attention in the literature [2] despite the findings of a large survey showing that even the most accurate model managed to detect only around half of the diabetic individuals [13]. Lastly, the behavior concerning precision and recall of interpretable and ensemble models as well as the effect of balancing the minority class have seldom been investigated on the same dataset.

The choice of model is dependent on the trade-off between the simplicity and accuracy of the method. A decision tree produces clear rules that can be used by doctors, for example, a glucose threshold, but a simple decision tree is extremely

sensitive regarding class imbalances and small variations in the data. In contrast to the decision trees, ensemble models reduce this variability since it is possible to use various trees trained on bootstrap data samples [6] for making predictions. XGBoost produces tree ensembles sequentially and focuses on the minimization of overfitting.

This work addresses these issues by developing and evaluating a machine learning-driven autonomous diabetes diagnosis system on 27,551 anonymized patient records from the publicly available Kaggle Diabetes Prediction Dataset [16]. Its objectives are to build a diagnosis pipeline that improves diagnostic support for medical practitioners, to measure the precision and reliability of the proposed models on held-out patient data, and to compare them with existing approaches. The system cleans and encodes the records, standardizes the attributes, and trains a Decision Tree and an XGBoost classifier on a 70:30 split, returning each prediction as a binary diagnosis together with a probability score. XGBoost achieved an overall accuracy level of 96.7% with a precision of 93.5% on a test dataset composed of 8,266 records of which 9.1% were patients suffering from diabetes. The Decision Tree showed a higher recall but resulted in six times more false positives. A random forest experiment based on sampling of the minority class indicates that substantial sensitive information can be retrieved if class imbalance is explicitly addressed.

The paper is structured in the following way. In the Related Work section, machine learning solutions to diabetes diagnosis are reviewed. In the Proposed Methodology section, the design of the system, preprocessing and feature extraction steps, and description of the classifiers used are introduced. The Experimental Setup section explains the data partitioning, implementation and evaluation metrics. Finally, in the Results and Discussion section, the confusion matrix and performance measures are presented, the results are compared with previous investigations, limitations are discussed, and the conclusion and the thoughts on future investigations are offered.

The major contributions of this work are as follows:

- A complete autonomic diagnostic process including data acquisition, preprocessing, partitioning, training and validation of the predictive model and producing a diagnosis with a probability score for clinical decision-making.
- An evaluation on 27,551 anonymized patient records from the publicly available Kaggle Diabetes Prediction Dataset, described by eight routinely available clinical and laboratory indicators, which allows the experiments to be reproduced by other researchers.
- A controlled comparison of an interpretable Decision Tree and an XGBoost classifier under the same 70:30 protocol, reported with confusion matrices, precision, recall, specificity and F1-score so that performance on the minority diabetic class is visible.
- An analysis of the precision–recall trade-off, showing that XGBoost reduces false-positive diagnoses from 217 to 36, complemented by a random forest experiment in which SMOTE-based rebalancing raises recall to 83.3%.
- A feature importance analysis confirming that HbA1c and blood glucose dominate the predictions, followed by age and body mass index.

Related Work

It has been observed that machine learning reviews for disease diagnosis give an implementation context. In a study by Ahsan et al. [2] that reviewed 102 studies concluded that image diagnostics are mainly affected by the presence of convolutional neural networks, on the contrary, structured clinical databases mainly utilize SVM, regression analysis, random forest, decision tree and neural networks as the most widespread approaches, more importantly, it has been concluded that hybrid and ensemble models tend to achieve better results in diagnostic compared to other single models. The same review pointed to three important gaps that have not received attention. First of all, class imbalance has never been reported. Second, almost all models studied lack

model interpretation. Finally, results obtained in one setting were hardly ever validated in another setting.

Deep learning methods have demonstrated their highest accomplishments with unstructured data. Gulshan et al. [4] trained a deep convolutional neural network on retinal images and correctly identified diabetic retinopathy with high accuracy. Similarly, Hannun et al. [5] developed a system for arrhythmias detection achieving results comparable with cardiologists. Most importantly, these applications show that results produced by computers are as good as the results made by doctors but rely on large sets of annotated data. On the other hand, the standard methodology for diabetes screening is much simpler, relying on a small number of medical and demographic parameters whereby tree-based ensembles outperform deep learning networks in terms of performance.

Most studies on structured data have compared classical classifiers on the Pima Indians Diabetes dataset. Farajollahi et al. [8] evaluated logistic regression, decision tree, SVM, XGBoost, random forest and AdaBoost on predictors such as pregnancies, BMI, insulin and age; AdaBoost achieved the highest accuracy, at 83%, but differences among the stronger models were small. Kırğıl et al. [9] tested eight classifiers, including J48, naive Bayes, SVM, logistic regression, multilayer perceptron, k-nearest neighbors, logistic model trees and random forest, after feature selection, imputation and normalization, and obtained a best accuracy of 80.87% with random forest. Olimboyeva et al. [11] applied naive Bayes and a linear-kernel SVM to standardized features and assessed them with confusion matrices, F1-measure and AUC-ROC, emphasizing that simple, well-calibrated models remain useful for preliminary screening. Although these studies are simple to execute and reproduce, they reach a maximum limit of around 80-85% accuracy. Given their use of a single dataset from backstock populations, these findings offer limited information regarding the performance of their models on patients.

Different ensemble designs have been suggested to overcome this plateau. For example, Rashid et al. [10] created a voting ensemble combining decision tree methods; logistic regression; k nearest neighbors; random forests; XGBoost; subjecting the combined model to standardization; imputations and local outlier factor filtering and achieved an accuracy of 81% using the same data. The results demonstrate that although combining different learners results in better stability; it cannot solve the issues arising from insufficient or noisy data. There are two most widely applied types of ensembles; the first one includes random forests; which reduce variance through averaging the no-correlated trees created from bootstrap samples[6], while the second one uses sequential tree-building procedures for correcting residual errors and enriching the XGBoost process with L1/L2 regularization methods; shrinkage; through column subsampling [7]. The latter properties are directly relevant to laboratory data, in which some tests are not ordered for every patient.

Larger and population-specific datasets give a more realistic picture of diagnostic performance. Xie et al. [13] trained eight models on 138,146 participants from the 2014 Behavioral Risk Factor Surveillance System, of whom 20,467 had type 2 diabetes; the neural network achieved the highest AUC (0.795) and accuracy (82.4%), but the decision tree achieved the highest sensitivity, at only 51.6%, and was recommended for initial screening. Granillo and Goldsztein [12] used logistic regression on survey-based health indicators and stressed the value of transparent models for physician trust. The research in South Asia carried out by Islam et al. [14] has shown the efficacy of the random forest algorithm as compared to naive Bayes and logistic regression, with the former boasting better performance than the two algorithms. The researchers collected score records for the data from Sylhet Diabetes Hospital, creating a self-assessment tool. Similarly, Prama et al. [15] reported about DiaBD dataset consisting of 5,288 records from Bangladeshi population consisting of glucose, BMI, blood pressure and family history variables with a low percentage of diabetic records (6.5% of total number of records).

The accuracy of random forest when applied to DiaBD was almost total (99.26%), however, it was about 85% when it comes to Pima data. The differences in the accuracy scores illustrate how important the population and class balance are for the performance results obtained.

Summing the literature evidence up, it should be noted that quite a gap is present in research in the area. While survey/questionnaire data approaches yield only moderate accuracy as reported by research [13], [14], it is clear from the previous information that Pima-based approaches can achieve the accuracy level of about 80%–85% [8], [9], [10], while studies involving direct glucose measures, such as DiaBD [15], are just emerging. In addition, most studies still emphasize accuracy, despite repeated evidence that class imbalance and interpretability determine whether a model is usable in practice [2]. What is needed is a diagnosis system trained on routine clinical and laboratory indicators that are widely available, comparing an interpretable model with a regularized ensemble under one protocol, reporting precision, recall and F1-score on the minority diabetic class alongside accuracy, and examining how far explicit rebalancing improves sensitivity. The system proposed in this paper is designed to meet these requirements.

Proposed Methodology

The proposed system is a supervised, autonomous diagnosis framework that assigns each patient to a diabetic or non-diabetic class from routinely recorded clinical and laboratory indicators and can report the probability associated with that decision. It is organized as a sequential pipeline of six layers, shown in Figure 1: data collection, data preprocessing, data splitting, a learning module for model training and testing, model evaluation, and diagnosis output. Every layer must perform just function, allowing any layer to be substituted for another layer, for instance, by incorporating a new algorithm in the learning module, without having to change anything else in the process.

Initially, the patient reference comes to the data collection layer, where the data are stored into structured files. Preprocessing layer cleans up the reference data and converts them into a uniform numerical form, after which the splitting layer separates the data into the training and testing sets. The learning module trains a Decision Tree and XGBoost algorithm on the training set and tests them on the testing set, and compares their predictions against the initial ones. The evaluation layer summarizes this comparison through confusion matrices and classification metrics and the prediction layer gives a binary diagnosis and probability of diabetes for every new patient.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

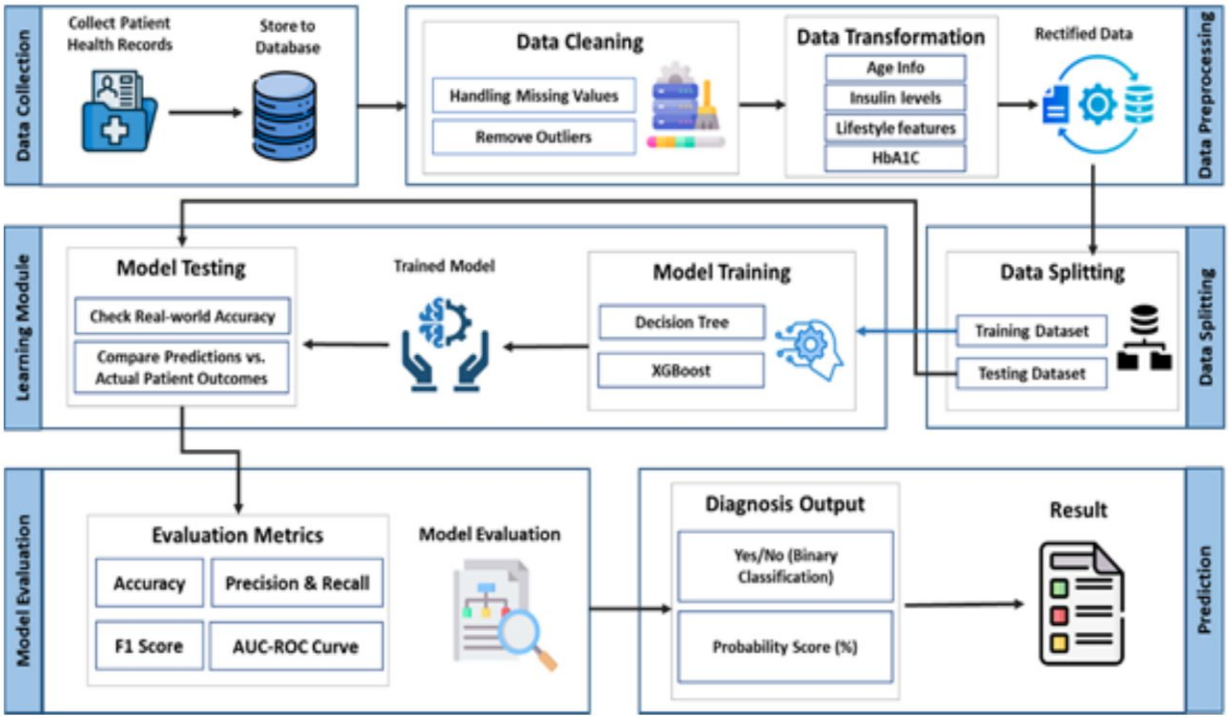


Figure 1: Proposed Machine Learning-Driven Autonomous Diabetes Diagnosis System

A. Dataset

The Diabetes Prediction Dataset is used for the experiments, which is a public dataset containing anonymized patient data available in Kaggle [16]. This dataset contains eight features plus a binary outcome determining diabetes state of the patient, as summarized in table 1. Two features belong to the area of real-time glucose measurement, namely HbA1c and blood glucose. Other attributes are demographic data and health state indicators, which also include diseases and smoking history of the patient. The experiment involved a subset of 27,551 records out of the whole data set after cleaning the dataset.

Table 1: Attributes of the Dataset

Attribute	Type	Description
Gender	Categorical	Female or male
Age	Numerical	Age of the patient in years
Hypertension	Binary	Presence (1) or absence (0) of hypertension
Heart disease	Binary	Presence (1) or absence (0) of heart disease
Smoking history	Categorical	Recorded smoking status
BMI	Numerical	Body mass index (kg/m²)
HbA1c level	Numerical	Glycated hemoglobin (%)
Blood glucose level	Numerical	Blood glucose concentration (mg/dL)
Diabetes (target)	Binary	Diabetic (1) or non-diabetic (0)

B. Data Preprocessing

Duplicates, missing values and rare categories were checked in the raw file. It had no missing values; thus, no imputation was necessary. A total of 336 duplicate records were deleted, and records in the - other gender category- were eliminated (this constitutes a negligible fraction of the data) in order to treat gender as a binary attribute. Categories in which smoking had been originally recorded were collapsed into coarser categories (current, ex-smoker, and none) to reduce sparsity. Categorical attributes were converted to integer codes with label encoding. All attributes were then standardized with the z-score transformation,

$$z = (x - \mu) / \sigma$$

where x is the original value and μ and σ are the mean and standard deviation of the attribute estimated on the training partition only; the same parameters were then applied to the testing partition, so that no information from the test data entered model fitting. Tree-based models don't require standardization, but it was kept as it ensures uniform representation for all classifiers in the pipeline and also allows for the addition of other scale-sensitive algorithms without having to alter the preprocessing step.

C. Feature Analysis

Covariates included all eight attributes and assessed their relevance in exploratory analysis, correlation matrix with the outcome, and impurity-based feature importances of a tree ensemble. Hence, the analyses showed that HbA1c and blood glucose were great predictors of the outcome, followed by age and BMI, with gender and smoking history playing minor roles: a finding that is concordant with clinical knowledge and indicates that the models base their decisions on physiologically meaningful indicators.

D. Decision Tree Classifier

The decision tree is a non-parametric classifier that partitions the given feature space recursively, using an appropriate threshold for each of the individual attributes. The partitioning at each node is set so as to maximize the reduction in impurity, measured with the Gini index:

$$Gini(D) = 1 - \Sigma_i \, p_i^2$$

where p_i indicates the fraction of observations in node D that belong to class i. The algorithm stops allocating any further splits if any of the nodes turns out to be pure/ or if no further split results in any further reduction of impurity. The resulting decision tree can be interpreted as a set of if-then rules, where certain blood glucose level should be higher than a certain HbA1C level in order to conclude that a patient is diseased. This transparency is the main reason for including the Decision Tree, although a single fully grown tree is sensitive to class imbalance and to small perturbations of the training data.

E. XGBoost Classifier

XGBoost builds an additive ensemble of trees in which each new tree is fitted to the gradient of the loss with respect to the current ensemble prediction [7]. Training minimizes a regularized objective:

$$Obj = \Sigma_i \, l(y_i, \hat{y}_i) + \Sigma_k \, \Omega(f_k), \quad \Omega(f) = \gamma T + \tfrac{1}{2} \lambda \, \Sigma_j \, w_j^2$$

where l is the logistic loss between the true label y_i and the prediction $\hat{y}_i$, f_k is the k-th tree, T is the number of leaves in a tree, w_j are the leaf weights, and γ and λ control the penalty on tree complexity. Shrinkage, the complexity penalty and sparsity-aware split finding limit overfitting and allow missing values to be handled natively [7], which makes XGBoost a suitable high-accuracy counterpart to the interpretable Decision Tree.

F. Random Forest with Class Rebalancing

To examine how much sensitivity can be recovered when the class imbalance is addressed explicitly, a complementary experiment was carried out with a random forest [6]. In this pipeline, the numerical therapy were standardized and categorical therapy one-hot encoded and balanced in the training data in two steps, the first of which involved using the synthetic minority over-sampling technique (SMOTE) to increase the diabetic class to 10% of the dominant class, int then random under sampling to reduce the majority class to 2:1 ratio. This was done in cross-validation pipeline where synthetic samples were generated only from training folds.

G. Diagnosis Output

The model generates the predicted class for each patient through the posterior probability of diabetes, considering a threshold of 0.5. The model's reporting of the probability rather than just the label allows decision making as intermediate probabilities can be used for further tests and the threshold can be adjusted for sensitivity or specificity, depending on clinical priority.

Experimental Setup

A. Dataset Partitioning

The experiments yield results of Decision Tree and XGBoost where the population of the study 27,551 records was separated into two sets of records in an arbitrary manner, with 19,285 records for training purposes (70% of total records) and 8,266 records used as a testing part (30% of total number of records). It is reported that there were 751 records of diabetic patients and 7,515 records of non-diabetic individuals, which makes a percentage of diabetic population 9.1%. It is worth noting that training sample and testing sample are taken from the same set of records without any bias in terms of class distributions (Table 2). The random forest experiment used a separate 80:20 split with 5,511 test records, of which 484 (8.8%) were diabetic.

Table 2: Data Partitions Used in the Experiments

Experiment	Partition	Records	Diabetic records	Diabetic (%)
Decision Tree, XGBoost	Training (70%)	19,285	–	–
Decision Tree, XGBoost	Testing (30%)	8,266	751	9.09
Random forest	Training (80%)	22,040	–	–
Random forest	Testing (20%)	5,511	484	8.78

B. Implementation

All experiments were implemented in Python in Google Colab. Pandas and NumPy were used for data handling, scikit-learn for preprocessing, the Decision Tree and random forest classifiers, grid search and evaluation, the XGBoost library for gradient boosting, imbalanced-learn for SMOTE and random under sampling, and Matplotlib and Seaborn for visualization. The Decision Tree and XGBoost classifiers were trained with the libraries' default hyper parameters (for XGBoost, 100 trees with a maximum depth of 6 and a learning rate of 0.3) and a fixed random seed. For the random forest, the number of trees (50, 100, 200), the maximum depth (unlimited, 10, 20), the minimum samples per split (2, 5, 10) and the minimum samples per leaf (1, 2, 4) were tuned by grid search with five-fold cross-validation on the training partition; the selected configuration used 100 trees, a maximum depth of 10, a minimum of five samples per split and four samples per leaf.

C. Evaluation Metrics

Because the data are imbalanced, accuracy alone would overstate performance: a model that labelled every patient as non-diabetic would still reach about 91% accuracy. Performance was therefore assessed from the confusion matrix, in which TP and TN denote correctly classified diabetic and non-diabetic patients and FP and FN denote non-diabetic patients labelled diabetic and diabetic patients missed by the model. The following measures were computed:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision = TP / (TP + FP)
Recall (Sensitivity) = TP / (TP + FN)
Specificity = TN / (TN + FP)
F1-score = 2 × (Precision × Recall) / (Precision + Recall)
Balanced Accuracy = (Recall + Specificity) / 2

Recall is the most important measure from a screening perspective, because a missed diabetic patient may go untreated, whereas precision

governs the number of unnecessary confirmatory tests. The F1-score and balanced accuracy summaries performance on both classes without being dominated by the majority class.

Results and Discussion

Table 3 and Figure 2 present the confusion matrices of the Decision Tree and XGBoost on the 8,266 test records. The Decision Tree correctly identified 547 of the 751 diabetic patients and 7,298 of the 7,515 non-diabetic patients, while XGBoost identified 518 diabetic and 7,479 non-diabetic patients. The two models therefore differ mainly in how they distribute their errors: the Decision Tree made 217 false-positive and 204 false-negative predictions, whereas XGBoost made only 36 false-positive but 233 false-negative predictions.

Table 3: Confusion Matrices on the Test Partitions

Model	Test records	TN	FP	FN	TP
Decision Tree	8,266	7,298	217	204	547
XGBoost	8,266	7,479	36	233	518
Random forest + SMOTE*	5,511	4,841	186	81	403

*Evaluated on a separate 80:20 split; not directly comparable with the other two models.

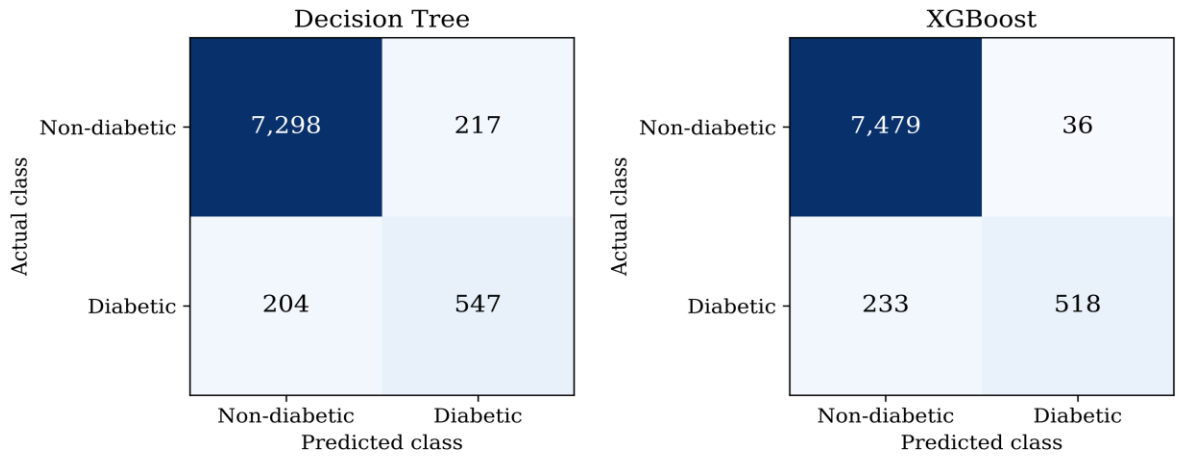


Figure 2: Confusion Matrices of the Decision Tree and XGBoost Classifiers on the 70:30 Test Partition

The metrics derived from these matrices are reported in Table 4 and compared visually in Figure 3. XGBoost achieved the higher accuracy (96.75% against 94.91%), precision (93.50% against 71.60%), specificity (99.52% against 97.11%) and F1-score (79.39% against 72.21%). The Decision Tree achieved the higher recall (72.84% against 68.97%), and the two models reached almost identical balanced accuracy (84.97% and 84.25%).

Table 4: Performance of the Classifiers on the Test Partitions (%)

Metric	Decision Tree	XGBoost	Random forest + SMOTE*
Accuracy	94.91	96.75	95.16
Precision	71.60	93.50	68.42

Metric	Decision Tree	XGBoost	Random forest + SMOTE*
Recall (Sensitivity)	72.84	68.97	83.26
Specificity	97.11	99.52	96.30
F1-score	72.21	79.39	75.12
Balanced Accuracy	84.97	84.25	89.78

*Separate 80:20 split with SMOTE and random under sampling applied to the training data.

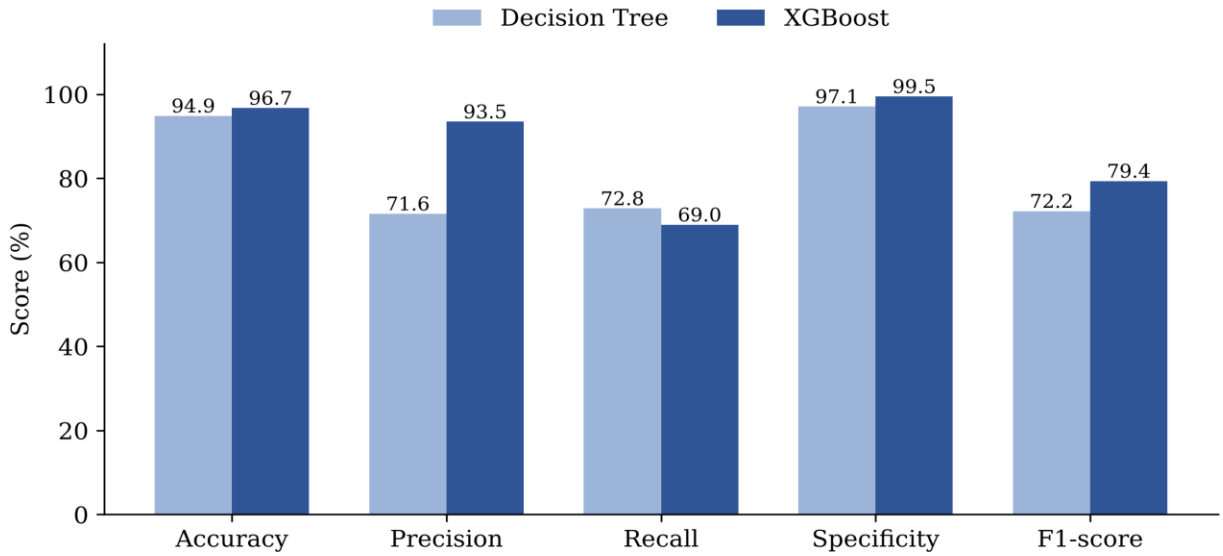


Figure 3: Comparison of Classification Metrics for the Decision Tree and XGBoost on the 70:30 Test Partition

The most significant difference between these two major models is the issue of accuracy. When using the XGBoost model, diabetic patients are correctly identified 93.5% of the time, as opposed to only 71.6% for the Decision Tree model. The rate of the false positive in the XGBoost model is six times smaller than that in the Decision Tree model, indicating that patients are less stressed out because of fewer false diagnosis. This reason shows that boosting technique works as it builds associated trees based on already existing models identified in the ensemble while also applying penalty for complexity of trees [7].

The cost of this conservatism is sensitivity. XGBoost missed 233 of the 751 diabetic patients, 29 more than the Decision Tree, so that its recall of 68.97% is the main limitation of the unbalanced models. With only about one diabetic record in eleven, the training objective is dominated by the majority class, and the 0.5 decision threshold favors the non-diabetic label for borderline cases whose HbA1c and glucose values lie close to the diagnostic cut-offs. The random forest experiment shows how much of this sensitivity can be recovered: with SMOTE oversampling and random under sampling of the training data, recall rose to 83.26% and balanced accuracy to 89.78%, at the cost of a lower precision of 68.42%. Although this result is at a different split and, therefore, cannot be compared directly with the two previous models, it indicates rebalance, and threshold modification of

predicted probabilities, as being the most successful option for enhancing the system's sensitivity.

Balanced accuracies of the Decision Tree and the XGBoost are similar, meaning that the higher level of XGBoost accuracy can be explained mostly by its improved behavior with the majority class. This finding supports the argument of Ahsan et al. [2] that class imbalance is often ignored in diagnostic research which focuses on reporting the accuracy most of the time and proves that it is important to provide recall, specificity, and F1-score along with accuracy. In terms of application, the models can be used complementarily, as XGBoost is more effective in cases where false alarms may be expensive, while rebalanced ensemble model can be used when missed cases are of greater importance, and the Decision Tree allows for having clear and transparent rules for the clinical practitioner.

The impurity-based feature importances of the fine-tuned random forest is illustrated in Figure 4. The major share of importance is taken by HbA1c (0.44) and blood glucose (0.31), the two characteristics accounting for 75% of the overall importance. Other relevant factors are age (0.13) and BMI (0.07), with hypertension and heart disease contributing negligibly and gender and previous smoking history having a very minor contribution as well. Thus, it can be concluded that the models are mostly based on the glycemic data obtained during clinical diagnosis, which lends credibility to them, but also means that their predictive capacity depends on the availability of these measurements.

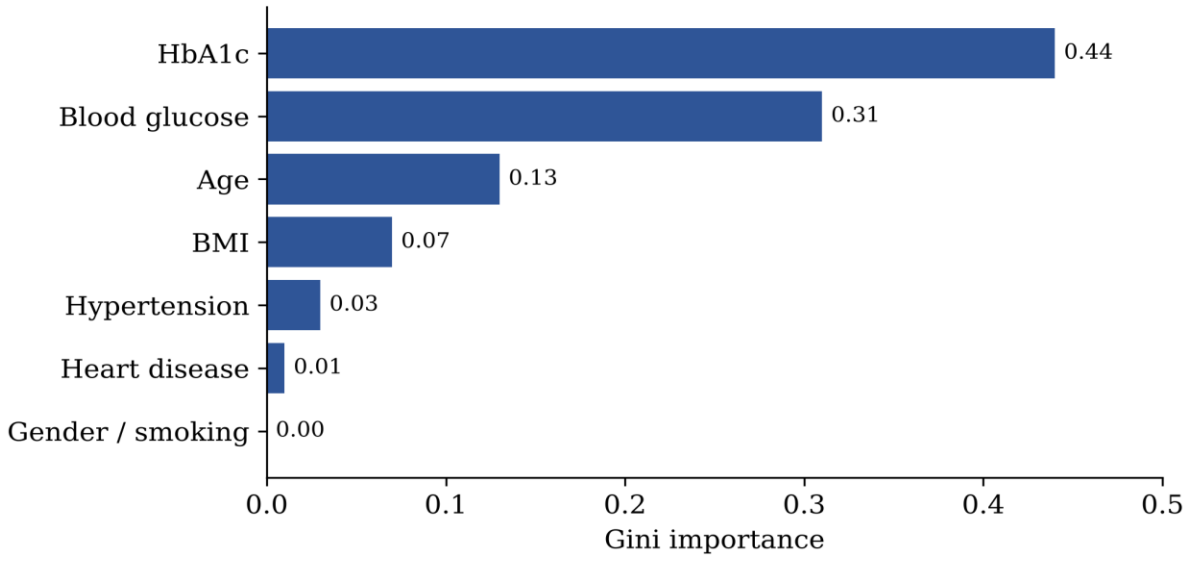


Figure 4: Feature Importances of the Random Forest Model

Table 5 places these results in the context of previously published studies. The comparison is indicative rather than head-to-head, because each study used a different dataset, population and attribute set.

Table 5: Comparison with Previously Published Diabetes Prediction Studies

Study	Dataset	Method	Reported Result
Farajollahi et al. [8]	Pima Indians Diabetes	LR, DT, SVM, XGBoost, RF, AdaBoost	AdaBoost: Acc. 83%

Study	Dataset	Method	Reported Result
Kırğıl et al. [9]	Pima Indians Diabetes	Eight classifiers incl. J48, NB, SVM, MLP, RF	RF: Acc. 80.87%
Rashid et al. [10]	Pima Indians Diabetes	Voting ensemble (DT, LR, KNN, RF, XGBoost)	Acc. 81%
Xie et al. [13]	BRFSS 2014 (n = 138,146)	Eight models incl. NN, DT, RF, SVM	NN: Acc. 82.4%, AUC 0.795
Prama et al. [15]	DiaBD, Bangladesh (n = 5,288)	Random forest	Acc. 99.26%
This study	Kaggle Diabetes Prediction (test n = 8,266)	XGBoost	Acc. 96.75%, Prec. 93.50%, Rec. 68.97%, F1 79.39%

When compared to models based on the Pima data sets, which achieved an accuracy of about 81–83%, as per research from multiple sources [8], [9], [10], and models established on surveys such as from Xcie et al. [13], the proposed methodology achieved a considerable increase in accuracy. The most important factor is the quality of features, as the data set used the methodology is composed of actual measurements of meaningful indicators such as HbA1c and blood glucose, which contain much more informative, diagnostic value than survey questions and other mere attributes of reference data sets, which has been verified by feature importance measurement. DiaBD research [15] includes glucose measurement and has a similarly low diabetes prevalence and still shows an even better accuracy, which leads to a conclusion about how highly outcomes depend on the data set used.

Several limitations qualify these findings. The data come from a single public dataset whose collection protocol and patient population are only partly documented, so external validation on clinical records from hospitals and laboratories, including South Asian populations with a high disease burden, is required before the results can be generalized. The Decision Tree and XGBoost were fitted with default hyperparameters and later were evaluated on a random split. To acquire a more reliable estimate, a stratified repeated cross-validation and confidence interval technique should be adopted, and finally a random forest is to be relater evaluation on the same split as other models. Finally, because HbA1c and blood glucose are themselves diagnostic criteria for diabetes, part of the predictive performance may reflect the recovery of these criteria; evaluating the system without the glycemic attributes, or against an independent clinical diagnosis, would clarify its added value for screening.

Conclusion and Future Work

Diabetes places a heavy and growing burden on health systems worldwide [1]. This study developed a machine learning-driven autonomous diabetes diagnosis system that processes routine clinical and laboratory indicators through a pipeline of data collection, preprocessing, data splitting, model training, evaluation and diagnosis output. An interpretable Decision Tree and an XGBoost classifier were trained and compared on 27,551 anonymized patient records from the public Kaggle Diabetes Prediction Dataset, and a rebalanced random forest was used to examine the effect of the class imbalance.

On a held-out test set of 8,266 records with a diabetic prevalence of 9.1%, XGBoost achieved an accuracy of 96.75%, a precision of 93.50%, a specificity of 99.52% and an F1-score of 79.39%, reducing false-positive diagnoses from 217 to 36 compared with the Decision Tree, which in turn achieved slightly higher recall. Rebalancing the training data with SMOTE and random under sampling raised the recall of a random forest to

83.26%, and HbA1c and blood glucose were the dominant predictors. These results show that tree-based models built on routine indicators can screen for diabetes accurately with very few false alarms, and that sensitivity on the minority class can be substantially improved by addressing the imbalance explicitly.

Future work will address the limitations identified above. All models will be tuned and evaluated on the same stratified splits with repeated cross-validation, confidence intervals and ROC and precision–recall analysis, and the decision threshold will be optimized for screening use. External validation on clinical records from local laboratories and hospitals, including Pakistani patient cohorts, is needed to establish generalizability. Explainable AI techniques, such as Shapley additive explanations, will be added to show clinicians which attributes drive each prediction. Finally, integrating the system with continuous glucose monitoring, wearable devices and electronic health records would move it from periodic screening towards continuous, risk-stratified monitoring of patients with or at risk of diabetes.

REFERENCES

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021.

[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-based disease diagnosis: A comprehensive review," Healthcare, vol. 10, no. 3, art. 541, 2022.

[3] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019.

[4] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[5] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, no. 1, pp. 65–69, 2019.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[8] B. Farajollahi, M. Mehmannavaz, H. Mehrjoo, F. Moghbeli, and M. J. Sayadi, "Diabetes diagnosis using machine learning," Frontiers in Health Informatics, vol. 10, no. 1, art. 65, 2021.

[9] E. N. H. Kırğıl, B. Erkal, and T. E. Ayyildiz, "Predicting diabetes using machine learning techniques," in Proc. 5th Int. Conf. Theoretical and Applied Computer Science and Engineering (ICTASCE), 2022, pp. 137–141.

[10] M. M. Rashid, O. M. Yaseen, R. R. Saeed, and M. T. Alasaady, "An improved ensemble machine learning approach for diabetes diagnosis," Pertanika Journal of Science and Technology, vol. 32, no. 3, pp. 1335–1350, 2024.

[11] D. Olimboyeva, D. Ziyadullaev, D. Mukhamedieva, K. Khujamkulova, M. Teshaboyev, and G. Ziyodullaeva, "Solving diabetes diagnosis problems using machine learning," BIO Web of Conferences, vol. 121, art. 01003, 2024.

[12] Y. Granillo and G. H. Goldsztein, "Machine learning as a tool to the diagnosis of diabetes," Journal of Student Research, vol. 11, no. 1, 2022.

[13] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," Preventing Chronic Disease, vol. 16, art. 190109, 2019.

[14] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis, Advances in Intelligent Systems and Computing, vol. 992. Singapore: Springer, 2020, pp. 113–125.

[15] T. T. Prama, M. J. Rahman, M. Zaman, F. Sarker, and K. A. Mamun, "DiaBD: A diabetes dataset for enhanced risk analysis and research in Bangladesh," Data in Brief, vol. 61, art. 111746, 2025.

[16] M. Mustafa, "Diabetes prediction dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

International Journal of Advanced Research