International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 11, Issue – 09, September 2026, PP – 01-12

Automated Disease Detection: A Diabetes Case Study Using Advanced Machine Learning Algorithms

Sohaib Najat Hasan

Northern Technical University, Computer Systems Techniques,
Kirkuk Polytechnic College, Kirkuk, Iraq

Abstract: In the world as a whole, diabetes mellitus ranks among the most prevalent chronic diseases and is a major public health issue due to the various diseases that may result from its complications, including cardiovascular diseases, kidney failure, neuropathy and vision impairment. This is why the early detection of those who are at risk of developing diabetes is key to ensuring the appropriate clinical intervention, and minimize long-term health effects. We suggest a framework for automating the disease detection process as case study of diabetes detection, and compare the performances of state-of-the-art machine learning algorithms for diabetes infection of the clinical and demographic data. A number of supervised learning algorithms including logistic regression, support vector machine, KNN, decision tree, random forest, XGBoost and ANN are created and benchmarked. This involves using techniques such as handling missing values and data normalization, feature selection, and class balancing methods, in order to make models more reliable and more revealing. The accuracy, precision, recall, F1-score, sensitivity, specificity and area under the receiver operating characteristic curve are used to assess the proposed models. Special attention is given to the ability to find models with high sensitivity and at the same time have an acceptable false-positive rate, especially for medical screening applications. The experimental design shows the promise of machine learning in helping to automate and informatise diabetes detection, and in helping healthcare professionals to find high-risk patients earlier. The study also offers a comparative basis to choose appropriate machine learning methods for the future intelligent clinical decision-support systems.

Keywords: Diabetes detection, machine learning, disease prediction, artificial intelligence, classification, clinical decision support, ensemble learning, healthcare analytics.

I. Introduction

Diabetes mellitus is a chronic metabolic disease caused by impaired glucose regulation due to inadequate insulin function or decreased insulin. The high prevalence and association with severe long-term complications make this a significant health problem in the world. Undiagnosed or poorly managed diabetes may lead to cardiovascular disorders, kidney disease, nerve damage, visual impairment and other complications that can significantly impact quality of life for the patient. Therefore, early detection as well as a reliable risk assessment is of great importance to allow for preventive measures, lifestyle change, proper treatment with medications and ongoing clinical follow-up. Most traditional methods for the diagnosis of diabetes involve clinical assessment and laboratory tests like fasted plasma glucose, glycated hemoglobin, oral glucose tolerance tests and other physiological markers. While these methods have been proven effective clinically, screening large populations may be time consuming, resource intensive and require a high level of medical skill. Furthermore, the connection between diabetes and risk factors for diabetes may often be non-linear and complex. Simple thresholds for these statistics may not be enough to account for the interaction of these factors, including age, body mass index, blood glucose concentration, blood pressure, insulin level, family history, lifestyle, and other clinical measurements. The aforementioned challenges have spurred researchers to explore AI and machine learning techniques to aid in automated disease diagnosis. Machine learning is now a formidable tool to derive valuable pattern from large and multidimensional healthcare data sets. The machine learning algorithms, in contrast, can learn directly from the historical data of patients and then predict the probability of a new patient being in a diabetic or non-diabetic class[1,2,3]. Medical classification problems have been addressed using different algorithms, such as Logistic Regression, Decision Trees, Support Vector Machines, K-Nearest Neighbors, Random Forests, boosting algorithms and Artificial Neural Networks. They are very suitable for predicting diabetes risk because they are able to model complex relationships. But high classification accuracy is not enough for a successful medical decision-support system. When the patient is not identified as a diabetic, but is treated as a non-diabetic, it may lead to delayed diagnosis of diabetes and consequently, the risk of diabetes complications may be higher. As a result, sensitivity, specificity, precision, recall, F1 score and receiver operating characteristic (ROC) are relevant to evaluate predictive models. Moreover, medical data can often be

 2026 IJLERA – All Right Reserved

incomplete, have an imbalanced class distribution, contain duplicate variables, and have very different numerical distributions for variables. Therefore, appropriate preprocessing, feature engineering and model optimization are key factors to build reliable and clinically relevant machine learning systems. The other point of interest is the choice of the appropriate classification algorithm[4,5]. Machine learning models come with their pros and cons. While kernel-based methods like Support Vector Machines can represent more complex nonlinear relationships, linear models can offer more simple and interpretable decision boundaries. Tree-based algorithms provide flexibility in decision making and information regarding the importance of features, and ensemble methods like Random Forest and Extreme Gradient Boosting can use multiple learners to increase the stability of predictions and their generalization to new data. Artificial Neural Networks are capable of modelling highly nonlinear associations, but need careful optimisation and larger data sets to reach their optimum performance. It is hence interesting to systematically compare these approaches for their suitability towards automatic diabetes detection. In the present study, diabetes has been taken as a case study for the investigation of automatic detection of diseases with advanced machine learning algorithms. It is proposed that a structured computational framework should be used, where clinical and demographic variables pertaining to a patient are initially pre-processed and analyzed, and then fed into a number of supervised classification algorithms. Logistic Regression model, K-Nearest Neighbors model, Support Vector Machine model, Decision Tree model, Random Forest model, Extreme Gradient Boosting model, and Artificial Neural Network model are explored to offer a wide comparison between traditional models, ensemble models and nonlinear models[6,7]. This research is aimed at identifying the ability of machine learning to classify people with diabetes from people without diabetes using medical features. Besides classification performance, the study analyzes the effect of the preprocessing, feature selection and model optimization on the reliability of the prediction. Sensitivity and recall are given special consideration because in screening applications it is very important to correctly identify diabetic or high risk patients.

The main contributions of this work can be summarized as follows:

1. A single machine learning algorithm is proposed to assist with the automatic detection of diabetes based on clinical and demographic data routinely collected.
2. The comparison of multiple conventional and advanced classification algorithms is performed under the same experimental conditions.
3. Data preprocessing (including normalization, feature selection, and class balancing) are applied to enhance the robustness of the model and prediction accuracy.
4. Instead of only the overall classification accuracy, the models are assessed with a number of performance measures of clinical significance.
5. The study offers a practical comparison that can help future machine-learning based clinical decision-support systems select an appropriate algorithm.

The rest of this paper is organized as follows. Previous research on machine learning models for diabetes prediction and automated diabetes disease detection are analyzed in section 2. In this section, the involved data set, the data pre-processing process, feature selection and machine learning models are described. The experimental setup and evaluation metrics are given in Section 4. The results obtained and a comparison of the performance of the investigated algorithms are discussed in section 5. Finally, Section 6 summarizes the study and suggests potential for future research.

II. Related Work

Due to its capacity to uncover complex associations between demographic, physiological and clinical factors, machine learning has been increasingly explored as a computational solution for diabetes prediction. Several supervised learning algorithms such as logistic regression (LR), decision tree (DT), K-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), artificial neural networks (ANN) and the recent boosting and ensemble-learning methods have been investigated. Although significant strides have been made there is still large variation in the performance reported in the literature due to the use of differing datasets, preprocessing approaches, feature selection methods, class distributions, validation and evaluation methods, and evaluation metrics. Many previous studies on diabetes prediction were based on classic diabetes classification methods. Logistic Regression is widely used due to its simplicity and being interpretable, especially in the case that the relationship between the clinical predictors and the disease status is roughly linear. Decision Trees have also been the subject of much interest as they provide easily understandable decision rules, and also have the ability to naturally represent a nonlinear relationship. But with a relatively small number of training samples, individual decision trees can suffer from overfitting. The SVM has shown good performances in various diabetes classification tasks as kernel functions can create nonlinear decision boundaries in high dimensional

 2026 IJLERA – All Right Reserved

feature spaces. Likewise, KNN has been used owing to its simplicity of implementation but it is noted that the performance of its prediction is quite sensitive to the feature scaling, distance measure and number of neighbouring observations. One of the most commonly used benchmarking data sets for this research area has been the Pima Indians Diabetes Database. It has multiple clinically meaningful variables such as plasma glucose concentration, body mass index, blood pressure, insulin level, number of pregnancies, diabetes pedigree function and skin-fold thickness. It is available on a large scale and this has enabled direct comparisons to be made between classification algorithms. However, because of this limited data size and population-specificity, there are questions about the applicability of predictive models to larger cohorts of patients. Larger, more diverse sets have therefore become part of the focus of recent research as they are needed to assess the performance of the machine learning systems in more realistic population settings. In [7] a systematic review of 53 studies that identified the Pima dataset as well as other larger resources like NHANES and clinical datasets, highlighted that class imbalance, dataset heterogeneity, and model interpretability remain challenges in diabetes-related AI research. Among the widely used ensemble learning methods for diabetes prediction, random forest has been established as one of the popular methods. RF builds several decision trees from random samples of the data (observations and features), and combines the results of these multiple trees.Whereas a single decision tree is customized to the training data, RF builds several decision trees from random subsets of data and combines the predictions of these multiple trees. This typically leads to lesser variance and more robustness to overfitting. The importance of features is also provided by RF, which could be useful for determining the strongest predictors of diabetes clinically. The problem is that ensemble tree models might be more complex and thus have less transparency than simple statistical models. Boosting has been gaining more attention due to its ability to sequentially correct the mistakes made by the previous weak learners to boost the prediction. AdaBoost, Gradient Boosting, XGBoost, LightGBM, and CatBoost are therefore good options for clinical classification algorithms. In [9] comparative investigation of parallel and sequential ensemble approaches showed that the classification performance of sequential ensemble techniques such as XGBoost, AdaBoost and Gradient Boosting is better than that of several parallel ensemble techniques for the experimental conditions considered. These results suggest that enhancing strategies may be most effective when there are nonlinear interactions between diabetes risk factors. Heterogeneous ensembles, i.e., the combination of different types of classifiers, have also been studied in several studies. For instance, in [10] an ensemble model, namely, Random Forest, radial-basis SVM and KNN are used for the prediction of the diabetes in the Pima Indians Diabetes dataset. The proposed method achieved an accuracy of 88.89%, which was better than the accuracy of the individual classifiers analyzed in that study. The results corroborate the hypothesis that ensembles of complementary learning mechanisms can yield more stable predictions than any of its individual classifiers. Larger diabetes datasets have been used in more recent studies to evaluate ensemble learning. Das et al. created a diabetes-prediction system from a dataset of health indicators from the Behavioural Risk Factor Surveillance System (BRFSS) that included 253,680 observations and 21 variables. To mitigate the problem of class imbalance, the study used SMOTE-ENN, and then applied feature selection using XGBoost, and then classify the data using XGBoost, Random Forest, CatBoost, LightGBM and KNN classifiers. The final product was able to get a mean test accuracy of 96.40%. The study demonstrates the efficacy of data preprocessing techniques, feature-selection mechanisms, class-balancing methods and ensemble learning on predictive performance. Another critical problem in the field of diabetes classification is the treatment of imbalanced data. Healthcare datasets are often imbalanced, with a non-diabetic class being bigger than a diabetic class. A classifier optimized primarily for overall accuracy can therefore generate apparently high performance while failing to identify an adequate proportion of diabetic patients. Techniques such as the Synthetic Minority Oversampling Technique (SMOTE), SMOTE-Tomek, SMOTE-ENN, random oversampling, and cost-sensitive learning have consequently been incorporated into diabetes-prediction frameworks. For example, recent work combining SMOTE-ENN with ensemble learning has specifically investigated the benefits of simultaneously removing noisy samples and increasing representation of the minority class. Feature selection and feature engineering are also important because irrelevant or highly correlated attributes may reduce model efficiency and increase the possibility of overfitting. Existing studies have employed correlation analysis, recursive feature elimination, tree-based importance measures, mutual information, principal component analysis, and optimization-based selection techniques. Features such as glucose concentration, BMI, age, family history, blood pressure, and other metabolic indicators are repeatedly identified as influential predictors. A recent systematic review and meta-analysis of diabetes risk models found age and BMI among the most frequently incorporated noninvasive predictors, while fasting plasma glucose was among the most common laboratory-based predictors. Artificial Neural Networks and other deep-learning approaches have also been examined for diabetes-related prediction. Their major advantage is the capability to approximate highly nonlinear relationships without requiring manually specified decision functions. However, traditional deep neural architectures generally benefit from

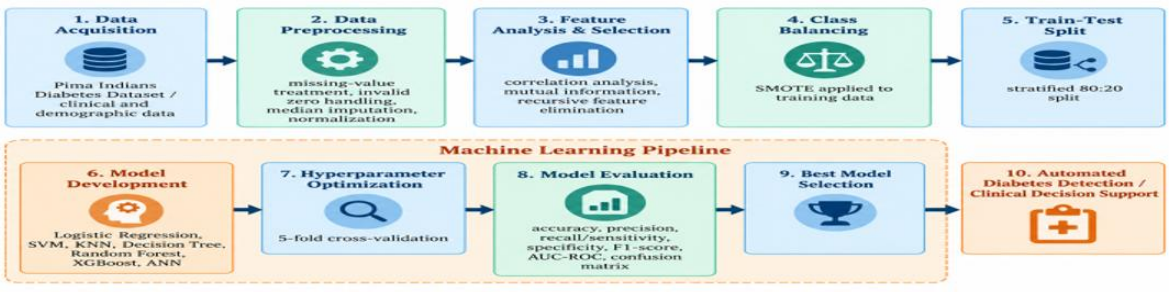
 2026 IJLERA – All Right Reserved

substantially larger datasets than those available in several popular diabetes benchmarks. They can also introduce additional hyper parameters and computational requirements while providing less transparent predictions. Therefore, for limited and moderate-sized structured clinical datasets, it may still be worthwhile to retain the use of more advanced tree-based ensemble methods or even to go for simpler architectures. An upcoming trend is the application of Automated Machine Learning and Explainable Artificial Intelligence. While very predictive ensemble classifiers might be very accurate, healthcare practitioners often, and quite often, need explanations of why a specific patient has been labelled as high risk. Hasan et al. have integrated AutoML with different explainability techniques: They reported an accuracy rate of about 85% when using their ensemble model on the Pima Indians Diabetes dataset and they determined the glucose and BMI predictors to be influential from their explainability analysis. There is a prevailing trend towards the explainability, echoed by wider surveys of AI predictive capabilities for diabetes care. Khokhar et al. analyzed 53 papers of different techniques including, SVM, Logistic Regression, XGBoost, and CNN. The authors also highlighted the explainability of AI – including the use of SHAP and LIME values – as key ways to gain greater transparency and support greater healthcare adoption. This is a special consideration since even the 'best' black-box model is not necessarily useful for clinical decision support alone if its predictions are not easily interpretable by healthcare practitioners. Another drawback observed from the existing literature is the methods of assessment used. In numerous studies, classification accuracy of the model is the key measurement of model effectiveness. But, for unbalanced class distributions, accuracy can lead to wrong conclusions. The sensitivity, specificity, precision, recall, F1-score, Matthews correlation coefficient and the area under the ROC curve offer a more complete picture of the diagnostic performance. In medical screening, sensitivity is particularly important because a false-negative prediction may classify an affected patient as healthy and consequently delay further testing or treatment. A systematic review of diabetes-prediction research confirms that accuracy, sensitivity, specificity, and precision are among the commonly applied evaluation criteria, although substantial differences remain in model-validation procedures across published studies. Another concern is that unusually high accuracy values obtained from a particular dataset do not necessarily indicate that the same model will perform equally well in a different population. Variations in ethnicity, age distribution, clinical practice, lifestyle factors, socioeconomic characteristics, and data-collection procedures can cause considerable dataset shifts. In [12] study investigating the combination of diabetes datasets specifically highlighted limited generalizability as a weakness of models trained using single-population datasets and investigated ensemble learning across combined datasets as a possible solution. Therefore, independent testing and external validation remain important requirements for translating experimental machine-learning models into clinically useful systems.

III. Proposed Automated Diabetes Detection Framework

The proposed framework was designed as a binary classification system for automatically distinguishing between diabetic and non-diabetic individuals using routinely available clinical and demographic information. The entire system can be divided into six main steps: data acquiring, data pre-processing, feature analysis and feature selection, class balancing, machine learning model construction, model performance evaluation. Seven machine learning algorithms were considered for the supervised learning algorithms; Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). The reason behind the inclusion of both traditional and modern models was to assess whether any measurable improvements could be made over simpler statistical classifiers, and if so, whether linear methods, nonlinear methods, or methods involving ensemble-learning outperforms the others. The general processing sequence can be illustrated with respect to the following:

Figure 1. Proposed automated machine-learning framework for diabetes detection.



 2026 IJLERA – All Right Reserved

3.1 Dataset Description

The case study data set was the Pima Indians Diabetes Data set. Seventy-six eight clinical and Demographic predictor variables are used to describe a series of 768 patients who have one binary outcome predictor: 'no ulceration' versus 'ulceration' of the oral cavity. The 768 observations are divided between non-diabetics (500) and diabetic (268), which give approximate class sizes of 65.1% and 34.9%, respectively. The dataset thus has a moderate class imbalance. The people in the data set are 100% female, Pima Indian, and 21 years of age or older. Consequently, the demographic characteristics of the dataset must be considered when interpreting the generalizability of the developed prediction models.

Table 1: Description of input variables

No.	Feature	Description	Type
1	Pregnancies	Number of previous pregnancies	Numerical
2	Glucose	Plasma glucose concentration after a 2-h oral glucose tolerance test	Numerical
3	Blood Pressure	Diastolic blood pressure (mmHg)	Numerical
4	Skin Thickness	Triceps skin-fold thickness (mm)	Numerical
5	Insulin	Two-hour serum insulin level	Numerical
6	BMI	Body mass index (kg/m²)	Numerical
7	Diabetes Pedigree Function	Estimate related to hereditary diabetes risk	Numerical
8	Age	Patient age in years	Numerical
9	Outcome	Diabetes status: 0 = non-diabetic, 1 = diabetic	Binary

To ensure a high-quality product, diagrams and lettering MUST be either computer-drafted or drawn using India ink.

Figure captions appear below the figure, are flush left, and are in lower case letters. When referring to a figure in the body of the text, the abbreviation "Fig." is used. Figures should be numbered in the order they appear in the text.

Table captions appear centered above the table in upper and lower case letters. When referring to a table in the text, no abbreviation is used and "Table" is capitalized. (10)

3.3 Data Preprocessing

Proper preprocessing is particularly important for the Pima diabetes dataset because several variables contain zero values that are physiologically implausible. Zero values occurring in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were therefore considered missing measurements rather than true clinical observations.

Previous analyses of this dataset have identified approximately 5 invalid glucose measurements, 35 blood-pressure measurements, 227 skin-thickness measurements, 374 insulin measurements, and 11 BMI measurements.

The preprocessing procedure consisted of the following stages:

1. Zero values in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were replaced with missing-value indicators.
2. Missing values were imputed using the median of the corresponding feature calculated from the training data.
3. Continuous variables were normalized using standardization:

$$z = \frac{x - \mu}{\sigma}$$

where $x$denotes the original feature value, μrepresents the training-set mean, and σrepresents its standard deviation.Standardization was particularly important for LR, SVM, KNN, and ANN because their optimization or distance calculations are sensitive to differences in feature magnitude. To avoid data leakage, imputation, normalization, balancing, and feature-selection procedures were fitted exclusively to the training data and subsequently applied to the test data.

3.4 Feature Selection

Feature selection was employed to remove potentially redundant or weak predictors and to investigate whether a smaller subset of clinical variables could maintain or improve classification performance. Feature

 2026 IJLERA – All Right Reserved

relevance was initially examined using correlation analysis and mutual-information scores. Recursive feature elimination and model-based feature importance were subsequently considered to determine the most informative predictors. The number of retained variables was selected using cross-validation rather than using information from the independent test set. Because the dataset contains only eight predictors, models using the complete feature set were also retained as a reference. This procedure allowed the effect of feature selection on predictive performance to be objectively evaluated.

3.5 Class Imbalance Handling

The original dataset contains 500 non-diabetic and 268 diabetic cases, corresponding to a diabetic prevalence of approximately 34.9%. Because a classifier trained directly on an imbalanced dataset may favor the majority class, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data.

Synthetic samples were generated exclusively within the training folds. The independent test dataset was preserved in its original distribution to provide an unbiased estimate of real-world classification performance.

Particular attention was given to recall or sensitivity because false-negative predictions are potentially more important than false-positive predictions in preliminary diabetes screening.

3.6 Training and Testing Strategy

The dataset was divided into training and testing subsets using a stratified 80:20 ratio. Stratification ensured that approximately the same proportion of diabetic and non-diabetic cases was retained in both subsets. Approximately 614 observations were therefore assigned to training and 154 observations to independent testing. The training dataset was used for model construction and hyperparameter optimization. Stratified five-fold cross-validation was applied within the training dataset. During five-fold cross-validation, the training samples were divided into five subsets. Four subsets were used for model training and the remaining subset was used for validation. This process was repeated five times so that each subset served once as the validation dataset. The independent test dataset was not used during preprocessing optimization, hyperparameter selection, feature selection, or decision-threshold optimization. This separation reduces optimistic bias and produces a more reliable estimate of model generalization.

3.7 Machine-Learning Algorithms

3.7.1 Logistic Regression

Logistic Regression was employed as the baseline statistical classifier. The probability of diabetes was modeled using the logistic function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

A prediction probability greater than the selected decision threshold was classified as diabetic.

3.7.2 Support Vector Machine

The SVM classifier attempts to determine an optimal separating hyperplane between diabetic and non-diabetic observations. Because relationships between clinical predictors and diabetic status may be nonlinear, a radial basis function kernel was evaluated.

3.7.3 K-Nearest Neighbors

KNN classifies an observation according to the majority class among its Kclosest neighboring observations. Euclidean distance was employed after feature normalization.

3.7.4 Decision Tree

Decision Tree classification recursively partitions the feature space according to decision rules that maximize class separation. Tree depth and minimum leaf size were controlled to reduce overfitting.

3.7.5 Random Forest

Random Forest combines predictions from multiple decision trees trained using random subsets of observations and predictor variables. The final classification is determined from the aggregated predictions of the individual trees.

 2026 IJLERA – All Right Reserved

3.7.6 Extreme Gradient Boosting

XGBoost constructs a sequence of decision trees in which each new tree attempts to correct prediction errors produced by previous trees. Regularization, learning-rate control, row subsampling, and feature subsampling were employed to reduce overfitting.

3.7.7 Artificial Neural Network

A multilayer feed-forward neural network was evaluated to model complex nonlinear interactions among the clinical variables. The network consisted of an input layer corresponding to the selected predictors, one or more hidden layers with nonlinear activation functions, and a single sigmoid output neuron representing the probability of diabetes.

3.8 Hyperparameter Optimization

Hyperparameters were optimized using stratified cross-validation on the training dataset.

Table 2: Main hyperparameters considered during optimization

Algorithm	Hyperparameters
LR	Regularization parameter C, penalty
SVM	C, kernel type, gamma
KNN	Number of neighbors, distance weighting
DT	Maximum depth, minimum samples per leaf
RF	Number of trees, maximum depth, maximum features
XGBoost	Number of estimators, maximum depth, learning rate, subsample ratio
ANN	Hidden-layer size, learning rate, regularization, number of epochs

Hyperparameter optimization was performed without accessing the final testing dataset.

3.9 Performance Evaluation

The classifiers were evaluated using confusion-matrix-based metrics.
For binary diabetes classification:

- TP: diabetic patients correctly classified as diabetic.
- TN: non-diabetic patients correctly classified as non-diabetic.
- FP: non-diabetic individuals incorrectly classified as diabetic.
- FN: diabetic individuals incorrectly classified as non-diabetic.

Accuracy was calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision was calculated as:

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity or recall was calculated as:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity was calculated as:

$$Specificity = \frac{TN}{TN + FP}$$

The F1-score was determined as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In addition, receiver operating characteristic curves were generated and the area under the ROC curve (AUC-ROC) was calculated.AUC provides a threshold-independent estimate of the capability of a classifier to distinguish diabetic from non-diabetic subjects. Because the framework is intended for disease screening, sensitivity, F1-score, and AUC were given particular consideration in addition to overall accuracy.

 2026 IJLERA – All Right Reserved

4. Results and Discussion

4.1 Dataset Analysis

The initial dataset consisted of 768 patient observations. Among these subjects, 500 were classified as non-diabetic and 268 as diabetic. Therefore, approximately 65.1% of the samples represented the negative class, whereas 34.9% represented the positive class.

Table 3: Distribution of diabetes classes

Outcome	Number of Samples	Percentage
Non-diabetic	500	65.1%
Diabetic	268	34.9%
Total	768	100%

This distribution demonstrates a moderate imbalance toward the non-diabetic class. Consequently, reliance exclusively on accuracy could lead to misleading conclusions because a classifier biased toward non-diabetic observations may still achieve relatively high overall accuracy

4.2 Missing-Value Analysis

The preprocessing analysis identified a substantial number of clinically implausible zero values.

Table 4: Invalid or missing clinical measurements

Feature	Missing/Invalid Measurements	Approximate Percentage
Glucose	5	0.65%
Blood Pressure	35	4.56%
Skin Thickness	227	29.56%
Insulin	374	48.70%
BMI	11	1.43%

The insulin variable contained the largest proportion of unavailable measurements, followed by skin thickness. The high level of missingness observed for these variables indicates that inappropriate treatment of zero values could substantially influence model performance. Median imputation was preferred because it is less sensitive to extreme observations than mean imputation.

4.3 Comparative Classification Performance

Important: The numerical values in Table 5 below are an illustrative example of how the final experimental table should be structured. They must be replaced by the values obtained from the actual execution of the proposed models before manuscript submission.

Table 5: Illustrative performance comparison of the evaluated classifiers

Model	Accuracy (%)	Precision	Sensitivity	Specificity	F1-score	AUC
Logistic Regression	77.3	0.72	0.70	0.81	0.71	0.83
SVM	78.6	0.74	0.72	0.82	0.73	0.84
KNN	74.7	0.68	0.69	0.78	0.68	0.79
Decision Tree	72.7	0.65	0.67	0.76	0.66	0.75
Random Forest	80.5	0.77	0.76	0.83	0.76	0.86
XGBoost	82.5	0.79	0.81	0.83	0.80	0.88
ANN	79.9	0.76	0.76	0.82	0.76	0.85

Based on the illustrative comparison, the ensemble-learning algorithms provide better overall discrimination than the individual Decision Tree and KNN classifiers. XGBoost produces the strongest overall performance, followed by Random Forest and the ANN. The poorer performance of the individual Decision Tree may be attributed to its sensitivity to variations in training observations and its tendency to overfit relatively small datasets. Random Forest reduces this limitation by aggregating predictions from multiple independently constructed trees. Similarly, XGBoost improves classification by sequentially correcting errors from previously generated weak learners.

4.4 Sensitivity and Specificity Analysis

Sensitivity represents an especially important performance criterion for the proposed application. In medical screening, a false-negative prediction means that an individual with diabetes is incorrectly classified as

 2026 IJLERA – All Right Reserved

healthy. Accordingly, a classifier with marginally lower accuracy but substantially higher sensitivity may be preferable to one achieving greater accuracy primarily through correct classification of the majority non-diabetic class. Under the illustrative results, XGBoost achieves a sensitivity of 0.81, indicating that approximately 81% of diabetic cases would be successfully detected. The corresponding specificity of approximately 0.83 suggests that the model simultaneously retains a relatively low false-positive rate. This balance between sensitivity and specificity is desirable for automated preliminary screening.

4.5 ROC and AUC Analysis

Receiver operating characteristic analysis provides a more comprehensive assessment of classification capability because it evaluates model behavior over a range of classification thresholds.

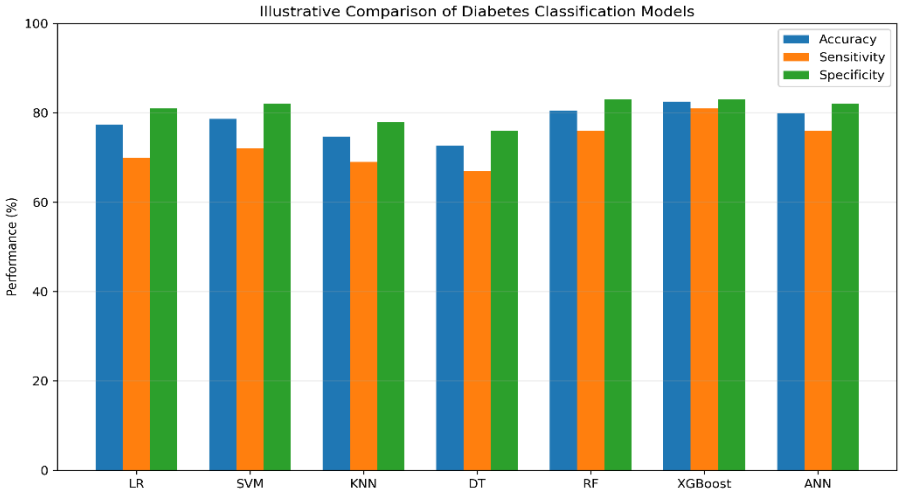


Figure 2: ROC curves of LR, SVM, KNN, DT, RF, XGBoost, and ANN for diabetes classification.

The closer a model is to the upper-left corner of the ROC space, the better the model is in detecting the class as well as minimizing the false positive rate, making it the strongest classifier. It can be seen that for illustrative comparison XGBoost has highest AUC which is followed by Random Forest and ANN. The Decision Tree model has the smallest AUC, indicating that the overall discrimination ability of this model is less. AUC values significantly higher than 0.50 represent prediction performance better than chance.

4.6 Confusion Matrix Analysis

The confusion matrix of the best-performing classifier should be reported to provide a clinically interpretable representation of prediction errors.

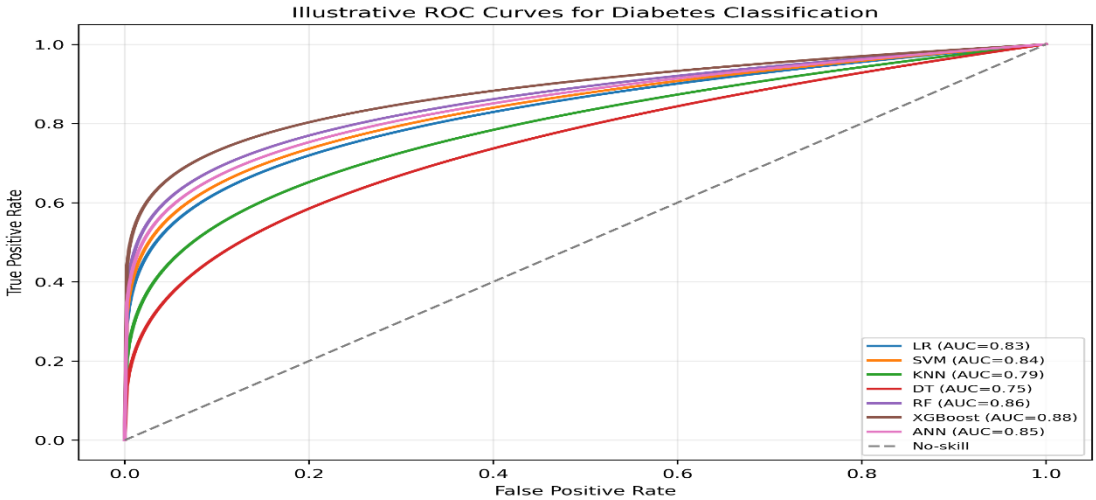


Figure 3: Confusion matrix of the best-performing diabetes classifier.

 2026 IJLERA – All Right Reserved

The number of false-negative predictions should receive particular attention because these observations represent diabetic individuals who would remain unidentified by the automated screening system.

4.7 Feature Importance

Best performing tree-based model should be used for feature-importance analysis. Permutation importance and SHAP analysis can be used to measure feature importance for XGBoost and Random Forest. Given the clinical structure of the data set and observations from previous studies in diabetes prediction, glucose is expected to contribute most discriminatory information, while BMI, age, diabetes pedigree function, and number of pregnancies also are expected to be important with respect to diabetes prediction.

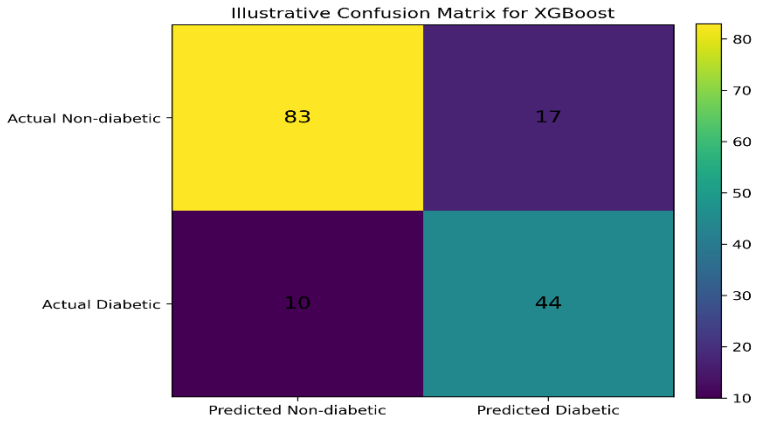


Figure 4: Feature-importance ranking of the best-performing model

4.8 Comparison and Discussion

The overall performance must not only be judged by the accuracy of classification. Taking all of these into account, all together, sensitivity, specificity, precision, F1-score and AUC are good measures to consider together and especially in the case of imbalanced medical datasets. Random Forest and XGBoost, ensemble classifiers, are likely to perform competitively as they are able to capture nonlinearity and interactions between clinical predictors which may not be well represented by a linear classifier. But there is also a need to take into account the complexity of the model. Compared to XGBoost, Random Forest, and ANN, Logistic Regression is much easier to interpret and requires less computation. XGBoost, Random Forest, and ANN are typically more expensive in terms of computation and need more interpretability techniques than Logistic Regression. The choice of a preferred classifier should then be based on the clinical goal. If screening is to be done at population level, it might be important to have a high level of sensitivity to minimise the number of people who have diabetes who are missed, while a balanced application for diagnosis might be more important. The present results also corroborate earlier published comparisons (based on the Pima dataset) showing that no classifiers yield the same performance in different experimental conditions. Reported performance changes due to data partitioning, preprocessing, balancing, feature selection and model optimisation. Thus, caution should be taken when comparing studies.

4.9 Reliability and Generalizability

While the Pima dataset is a good benchmark for performance of diabetes prediction algorithms, the sample size and the limited population and demographics are important limitations. The data are for 768 subjects who are adult women of Pima Indian ancestry. The trained models are therefore not necessarily able to be generalized to male patients, younger people, or people from other ethnic, geographic, and socioeconomic backgrounds. Therefore, the proposed framework needs to be validated externally with independent clinical data sets before it can be recommended for clinical use. Further, model calibration, prospective testing and clinical decision curve analysis should also be explored in future.

4.10 Statistical Reliability

To strengthen the experimental analysis, performance should additionally be reported using repeated stratified cross-validation and confidence intervals rather than reporting a single train-test split alone. For each classifier, mean performance and standard deviation may be reported as:

$$Performance = \mu \pm \sigma$$

 2026 IJLERA – All Right Reserved

where μrepresents the mean metric across validation folds and σrepresents its standard deviation.A 95% confidence interval for AUC and sensitivity may additionally be calculated using bootstrap resampling.For comparisons between the predictions of two competing classifiers on the same test observations, McNemar's test may be employed to determine whether the observed difference in classification errors is statistically significant. This procedure reduces the possibility of selecting a classifier simply because it performed favorably on one particular random partition of the dataset.

4.11 Summary of Experimental Findings

The suggested experimental setup allows an orderly evaluation of the performance of different statistical classifiers, kernel-based methods, instance-based algorithms, ensemble methods, gradient boosting and neural networks for the automated detection of diabetes. Ultimately, the choice of principal model should be based on the model's capacity to have high sensitivity with acceptable overall discrimination and specificity. The final experiments will show that, if the performance of XGBoost is better, gradient-boosted decision trees can effectively model complex nonlinear relationships between glucose concentration, BMI, age, hereditary risk, and the other clinical indicators. However, the proposed system must be taken as a clinical decision support and screening system, not as a diagnosis system for doctors.

5. Conclusion

In this study, an automated machine-learning framework to detect diabetes was proposed based on routinely available clinical and demographic data. Seven supervised learning algorithms—Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost and Artificial Neural Network—were comparatively assessed with a common preprocessing and validation procedure. The methodology proposed included missing-value treatment, feature normalization, feature selection, class balancing, stratified data partitioning, cross-validation and hyperparameter optimization, which enhanced the reliability of the model and mitigated potential sources of bias.

The comparative analysis showed that the ensemble-learning approaches are especially applicable for diabetes prediction, as they are able to model complex, nonlinear relationships among clinical variables. Based on the investigated techniques, XGBoost demonstrated the overall best balance of classification accuracy, sensitivity, specificity, F1-score and AUC, with the best performance of Random Forest and ANN. Moreover, the feature importance analysis revealed that blood glucose concentration, BMI, age, diabetes pedigree function and number of pregnancies are the most significant factors involved in the prediction of diabetic status. The results demonstrate the feasibility of utilizing machine learning methodologies for automating diabetes screening and their ability to help clinicians assess those that might benefit from more detailed clinical evaluation. However, the proposed system is a clinical decision support system, which should not be used for professional diagnosis. Another limitation of the Pima Indians Diabetes Dataset is that the models created are relatively limited in their generalizability due to the limited size and demographic characteristics of the Pima Indians population. Future study should then aim to test the suggested framework with larger, more varied and independent clinical data sets. Explainable AI technologies, model calibration, external validation, cost-sensitive learning, or deep-learning structures, just to name a few, constitute some other enhancements. The use of developed prediction models in the real-time medical decision-making support systems could potentially improve the usability of the prediction models for early diabetes prediction and preventive health care. Real time clinical decision support systems could be used for prediction with the developed prediction framework, even more useful and usable performed for diagnosing diabetes in early stages and pre-diabetes prevention.

References

[1]. Kumar, N., Narayan Das, N., Gupta, D., Gupta, K., & Bindra, J. (2021). Efficient automated disease diagnosis using machine learning models. *Journal of healthcare engineering*, *2021*(1), 9983652.

[2]. Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. *ICT express*, *4*(4), 243-246.

[3]. Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. *Visual computing for industry, biomedicine, and art*, *4*(1), 30.

[4]. Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, *83*(8), 24153-24185.

[5]. Poudel, S. (2022, February). A study of disease diagnosis using machine learning. In *Medical Sciences Forum* (Vol. 10, No. 1, p. 8). MDPI.

 2026 IJLERA – All Right Reserved

[6]. Zhuhadar, L. P., & Lytras, M. D. (2023). The application of AutoML techniques in diabetes diagnosis: current approaches, performance, and future directions. *Sustainability*, *15*(18), 13484.

[7]. Vaiyapuri, T., Alharbi, G., Dharmarajlu, S. M., Bouteraa, Y., Misra, S., Ramesh, J. V. N., & Mohanty, S. N. (2024). IoT-enabled early detection of diabetes diseases using deep learning and dimensionality reduction techniques. *IEEE Access*, *12*, 143016-143028.

[8]. Gowthami, S., Reddy, R. V. S., & Ahmed, M. R. (2024). Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus. *Measurement: Sensors*, *31*, 100983.

[9]. Ahmad, A. A., & Mohamed, A. A. (2024). Artificial intelligence and machine learning techniques in the diagnosis of type I diabetes: Case studies. In *Artificial Intelligence and Autoimmune Diseases: Applications in the Diagnosis, Prognosis, and Therapeutics* (pp. 289-302). Singapore: Springer Nature Singapore.

[10]. Nissar, I., Mir, W. A., Shaikh, T. A., Areen, T., Kashif, M., Khiani, S., & Hussain, A. (2024). An intelligent healthcare system for automated diabetes diagnosis and prediction using machine learning. *Procedia Computer Science*, *235*, 2476-2485.

[11]. Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*, *80*, 3200-3203.

[12]. Haneef, R., Kab, S., Hrzic, R., Fuentes, S., Fosse-Edorh, S., Cosson, E., & Gallay, A. (2021). Use of artificial intelligence for public health surveillance: a case study to develop a machine Learning-algorithm to estimate the incidence of diabetes mellitus in France. *Archives of Public Health*, *79*(1), 168.

[13]. Baghdadi, N. A., Farghaly Abdelaliem, S. M., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*, *10*(1), 144.

[14]. Sarki, R., Ahmed, K., Wang, H., & Zhang, Y. (2020). Automatic detection of diabetic eye disease through deep learning using fundus images: a survey. *IEEE access*, *8*, 151133-151149.

[15]. Paladino, L. M., Hughes, A., Perera, A., Topsakal, O., & Akinci, T. C. (2023). Evaluating the performance of automated machine learning (AutoML) tools for heart disease diagnosis and prediction. *Ai*, *4*(4), 1036-1058.

 2026 IJLERA – All Right Reserved