Empirical Evaluation of Different Algorithms to Assess The Probability of Diabetes in its Early Stages

Rania Ashraf[1*], Roz Nisha[2], Fahad Shamim[3], Shahzad Nasim[4], Sarmad Shams[5]

Abstract— **High blood sugar is a symptom of metabolic disorder, diabetes, an incurable and fatal disease. The primary cause of the disease is a hormone imbalance, which causes insulin impaction. Insulin is the specific hormone that regulates the sugar intake from the blood. The disease results in the body's inability to either make sufficient insulin or inadequate use of the produced insulin. Almost 1.6 million population die yearly due to this deadly disease. Early diagnosis can help reduce malignancy and enhance life expectancy. Since the medical data of diabetic individuals display a recognizable pattern, diabetes can be predicted in its early stages using machine learning algorithms. This is another way to get an early diagnosis without a glucose screening test. In this proposed paper, the prediction of early-stage diabetes is made by machine learning. The study individually experimented with eight machine learning algorithms over a dataset of 521 instances with 17 features. The performance assessment of every model is evaluated not only with accuracy metrics and confusion matrix, but AUC, F-score, recall, precision, TPR, & FPR are also observed to improve the algorithms' performance. The results of the applied techniques are validated using 5-fold cross-validation. AdaBoost classifier measures the lowest accuracy score with 82.89% accuracy. In comparison, the best score is measured by a Random Forest of 93.4. Similarly, the highest rating, calculated using Support Vector Machine, is 93.4 as well. Still, SVM exhibits a higher score of F-score and recall than RF, making it the best fit classifier for the study conducted. The rest of the classifiers have also performed well-having an accuracy of more than 80%. The findings indicate that the SVM Classifier is the most effective machine learning technique against binary-based classification datasets and can be utilized in predicting early-stage diabetes.**

Keywords— **Diabetes, Decision Tree, Naïve Bayes, Random Forest.**

INTRODUCTION

The term Diabetes mellitus is originated from the Greek word diabetes, which also means to siphon or pass through, and the Latin term Mellitus, meaning sweet. According to historical analysis, Apollonius of Memphis coined "diabetes" sometime between 250 and 300 BC [1]. The sweet character of the urine in this illness was discovered by the ancient Greek, Indian, and Egyptian civilizations, leading to the spread of the term diabetes mellitus [2]. In 1889, Mering and Minkowski discovered that the pancreas plays a part in the pathophysiology of diabetes. In 1922, Banting, Best, and Collip at the University of Toronto extracted the insulin hormone from the bovine pancreas, opening the door to creating a world-class diabetes treatment. Significant work has been done to address this expanding issue throughout the years, leading to several breakthroughs and the development of quality techniques. Regrettably, one of the most widespread chronic illnesses in the world is still diabetes. It is still Pakistan's seventh most frequent reason for demise. [3].

Diabetes, a chronic illness, is caused by the body's inability to generate adequate insulin, leading to a rise in the glucose level in the blood. The disease has the potential to destabilize global health. According to the studies conducted in 2019 by WHO, diabetes has entered the top 10 causes of death [4]. Findings of the current 10th edition of the International Diabetes Federation (IDF) confirm that one of the 21st century's most urgent worldwide health issues is diabetes. [5]. Approximately 6.7 million adults (20–79) are predicted to have passed away in 2021 due to diabetes or its consequences, and it is estimated that 537 million people have diabetes. By 2030 and 2045, this number is anticipated to reach 643 million and 783, respectively [5]. 90% of instances of diabetes are Type 2 diabetes. In comparison, the remaining 10% are primarily caused by Type 1 diabetes mellitus and gestational diabetes, while 374 million people are at risk of having type 2 diabetes due to the rising percentage [6]. Kidney failure, a higher risk of heart attack, and stroke are just a few severe health effects of diabetes. Uncontrolled diabetes can affect vital organs such as the eyes, kidneys, and the vascular system. There are various insulin types, namely, Type I and Type II.

Even though diabetes mellitus is lethal, early detection might lessen the risk if not treated in time. It is the only way to minimize the risk that can lead to complications if not treated early. For early diagnosis, numerous medical diagnostic techniques are already in use. Foreseeing diabetes is crucial for effective treatment to prevent the disease's subsequent consequences. Studies on disease classification, diagnosis, prediction, and treatment have been done in great numbers. Several Machine Learning algorithms have been used [7]–[9]. The algorithms can be used to identify and predict the disease. Promising findings from recent studies have been found in the risk prediction of diabetes mellitus [10]–[12]. But based on study data, hardly any of them was ever successful at achieving accuracy, that is, over 80% [13]. [14] showed the complete analysis of the on-hand research available on the classification model of diabetes prediction. Hence, a fast-processing system that can be used for prediction purposes to produce more accurate findings is needed. Therefore, a comparative study of a few classification techniques, including AdaBoost, Logistic Regression, and Random Forest, is the goal of this study to increase the accuracy by using a publicly available diabetes dataset.

On the diabetes prediction dataset, several classification methods have been applied to differentiate the cases into positive and negative groups. Utilizing different assessment criteria, such as classification accuracy, confusion matrix, AUC value, & F-score, their performances will be evaluated. Future studies can use the study's results as a guide to creating a baseline methodology for the best classification of diabetes mellitus.

[1-2-3-5] Liaquat University of Medical and Health Sciences, Jamshoro
[4] The Begum Nusrat Bhutto Women University, Sukkur
Country: Pakistan
Email: * raniya844@yahoo.com

When doing a background study on the related work of different researchers and authors, it is found that the feature that plays a crucial role and is selected as the ideal attribute can still not ensure 100% accuracy. Most researchers use several classification algorithms such as Bayesian Rule, SVM, decision tree, multilayer perceptron, KNN, and Logistic regression. In contrast, few studies employ recurrent neural networks or deep learning to anticipate cases accurately. Further research works considered for the comparative analysis are mentioned in Table I. The typical features which indicate the risk of early-stage diabetes and higher accuracy achieved by ensemble algorithms (especially Random Forest, i.e., 94%) are the conclusion of this paper.

The novelty of this study lies in its comprehensive comparison of eight machine learning algorithms, such as Random Forest (RF), Support Vector Machine (SVM), and AdaBoost, applied to early-stage diabetes prediction. It uniquely combines a wide set of performance metrics, including accuracy, recall, precision, F1-score, and AUC, to determine the most effective classifier. Additionally:

LITERATURE REVIEW

This systematic literature review evaluates the methodologies, datasets, algorithms, and performance metrics used in previous research to predict diabetes, emphasizing advancements and identifying gaps addressed by this paper

Table I: Comparative Literature Review

References	Study Purpose	Algorithms	Datasets	Evaluation Parameters	Key Findings
Chatrati 2022 [15]	To predict type-2 diabetes and hypertension	DT, LR, K-NN, DT SVM	PIDD	ACC, Scatter plot, ROC curve, CM,	SVM 75%
Ani 2016 [16]	To create a system for supporting clinical decisions for the prediction & prognosis of CRF.	10-fold cross-validation, ANN, KNN, DT, NB	UCI	ACC	DT 93%
Maniruzzaman 2020 [17]	Machine Learning based system to predict Diabetes	NB, AB, DT, the combination of RF-LR, NB	National Health and Nutrition Examination Survey	AUC, ACC	RF-LR 94.24%
Lynch 2017 [18]	To apply supervised learning techniques to classify the survival of lung cancer patients	DT, LR, GBM, SVM & RF	SEER database	RMSE	RF 15.63 (RMSE)
Kumari 2021 [13]	Increase the accuracy of diabetes prediction by different ML techniques	LR, NB, RF	PIMA diabetes & Breast carcinoma dataset	ACC, AUC, F1-score, precision, recall	97.02% on the breast carcinoma dataset 79.08% on PIMA dataset
Chen 2012 [19]	To predict the binding sites of microRNAs	SVM, NN, DT, RF	microRNA data	ACC	RF 75%
Rajendra 2021 [20]	Diabetes Prediction using LR and improved accuracy using various ensemble techniques	LR	PIMA & Vanderbilt	F1-score, precision, recall	78% for Dataset 1 93% for Dataset 2
Eskidere 2012 [21]	Parkinson's disease remote tracking using regression techniques	SVM, LS-SVM, GRNN, MLPNN	UCI archives	Mean absolute error	SVM 6.99 (Mean absolute error)
Yadav 2021 [22]	Using bagging & boosting classification techniques to predict diabetes	DT, JRIP, OneR, Bagging, Boosting Chi-Square for feature	UCI dataset	ACC, F1-score, precision, recall	Bagging Ensemble Method 98%
Chen 2013 [23]	Diagnosis of Parkinson's disease using fuzzy KNN	10-fold cross-validation, SVM, FKNN	UCI	ACC, ROC, AUC	RF 96.07%
Goyal 2022 [24]	Type-2 diabetes prediction using ensemble method and classification	10-folds cross-validation & ensemble method	PIDD	ACC	77.60%.

Prakash 2021 [25]	To increase the accuracy of early diabetes diagnosis	NBTree, RF, SimpleCART, and RandomTree	PIDD	ACC, precision, f-score, ROC, PRC & computational time	79.22%
Sadhu 2021 [26]	Study of different classifiers to predict early-stage diabetes with accuracy	DT, MLP, SVM, RF, LR, KNN, NB	UCI	ACC, f-score, ROC	RF 98.07%
Behroozi 2016 [27]	A framework of multiple classifiers for the detection of PD based on several vocal tests	KNN, SVM, DT, NB	UCI	ACC, specificity, ACC, sensitivity	SVM 87.50%
ERGÜN 2021 [28]	Early diabetes prediction with a machine learning method	10-fold cross-validation, DT, RF, KNN, XG boost	UCI	ACC, CM, precision, recall, f-score	CNN 99.04%
Saxena 2022 [29]	To study feature selection and classifiers to predict diabetes more accurately	MLP, DT, RF, KNN, feature selection techniques	PIMA	AUC, ACC	RF 79.8%
Tigga 2020 [30]	Utilizing incredibly accurate algorithms to forecast the likelihood of type 2 diabetes	RF, NB	PIDD	ACC, recall, precision, F1-score	RF 74.46% on both datasets
Jashwanth Reddy 2020 [31]	Diabetes analysis & early detection with machine learning	GB, NB, LR, SVM, KNN, RF	PIDD	ACC, precision, recall, ROC	RF 80%
Hussain 2018 [32]	Prostate cancer detection using machine learning and various feature-extracting strategies	DT, SVM, NB, k-fold validation	MRI Data	ROC, specificity, sensitivity, PPV, NPV, FPR	SVM 98.34%
Jackins 2021 [33]	To predict clinical disease, i.e., heart disease, diabetes, and cancer, with AI-based	RF, NB	PIDD	ACC	NB 74.64% RF 74.04%
Raghavendran 2022 [34]	Analyzing datasets to estimate the likelihood of type-2 diabetes using classification techniques	SVM, KNN, LR, DT, AB, RF, NB	PIDD	F1 Score, Precision, Recall, ACC, ROC & CM.	AB 95%
Laila 2022 [35]	To enhance the performance of algorithms by the Ensemble method	AB, Bagging, RF	UCI	F1 Score, Precision, Recall, ACC	RF 97%
Zupan 2000 [36]	A case study on survival prediction due to recurrence of prostate cancer	10-fold cross-validation, DT & NB	Clinical data	ACC, sensitivity, specificity	NB 70.8%
Hung 2017 [37]	Stoke prediction of the large-scale population by comparing DNN & other ML algorithms	ANN, LR, SVM	EMC database	ACC	ANN 0.873

PROCEDURE

A. Dataset

The dataset for creating the model is taken from the Google repository [38]. It contains 520 instances and 17 attributes. The dataset is summarized in Table II.

Dataset Analysis: The study uses a publicly available dataset, performs extensive preprocessing (removing duplicates, correlation analysis, and validation), and optimizes performance with 5-fold cross-validation.

Table II: Standard Feature Description

Attributes	Description
Age	20 years - 79 years
Sex	0. Male, 1. Female
Polyuria	1. Yes, 0. No.
Polydipsia	1. Yes, 0. No.
Sudden Weight loss	1. Yes, 0. No.
Weakness	1. Yes, 0. No.
Obesity	1. Yes, 0. No.
Polyphagia	1.Yes, 0. No.
Genital thrush	1. Yes, 0. No.
visual blurring	1. Yes, 0. No.
Itching	1. Yes, 0. No.
Irritability	1. Yes, 0. No.
Delayed Healing	1. Yes, 0. No.
Partial Paresis	1. Yes, 0. No.
Muscle Stiffness	1. Yes, 0. No.
Alopecia	1.Yes, 0. No.
Class	1. Positive, 0. Negative

The variables, positive or negative, are used to represent a diabetic and a non-diabetic patient, respectively.

B. Data preprocessing

Statistical data may have unclear, inaccurate, or random errors. No quality results are obtained if the input data quality is poor. To produce high-quality results, the data must be preprocessed. This is done by applying data cleaning, correlation analysis, and data splitting techniques to make the data more appropriate for using different diabetes risk prediction techniques.

1) Data cleaning

Data cleaning involves removing duplicate observations, fixing missing data, and validating data. In the data used, there were no null values present. However, eliminating duplications changed the number of instances from 520 to 251. The representation of attributes with their standards is given in Figure 1, Figure 2, and Figure 3. Data validation was done by answering the following questions:

- Does the data make complete sense now?
- Is the data reasonable and falls within a rational range?
- Is the data consistent with the working theory?

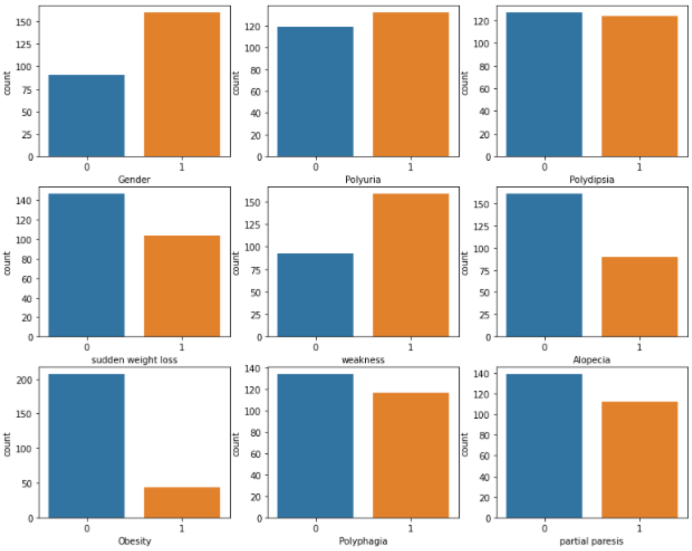


Figure 1: Representation of categorical attributes with their standards

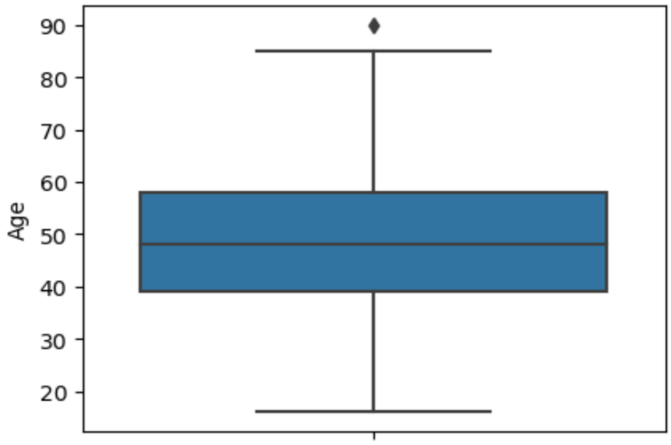


Figure 2: Representation of numerical attributes with their standards

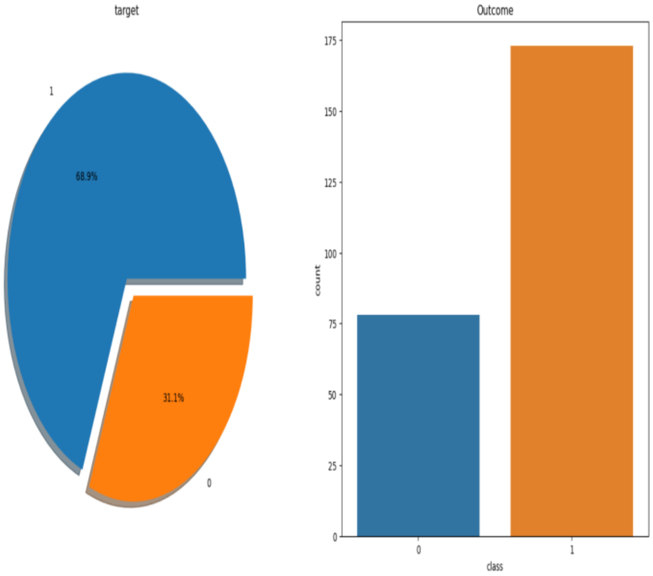


Figure 3: Representation of outcome with their standards

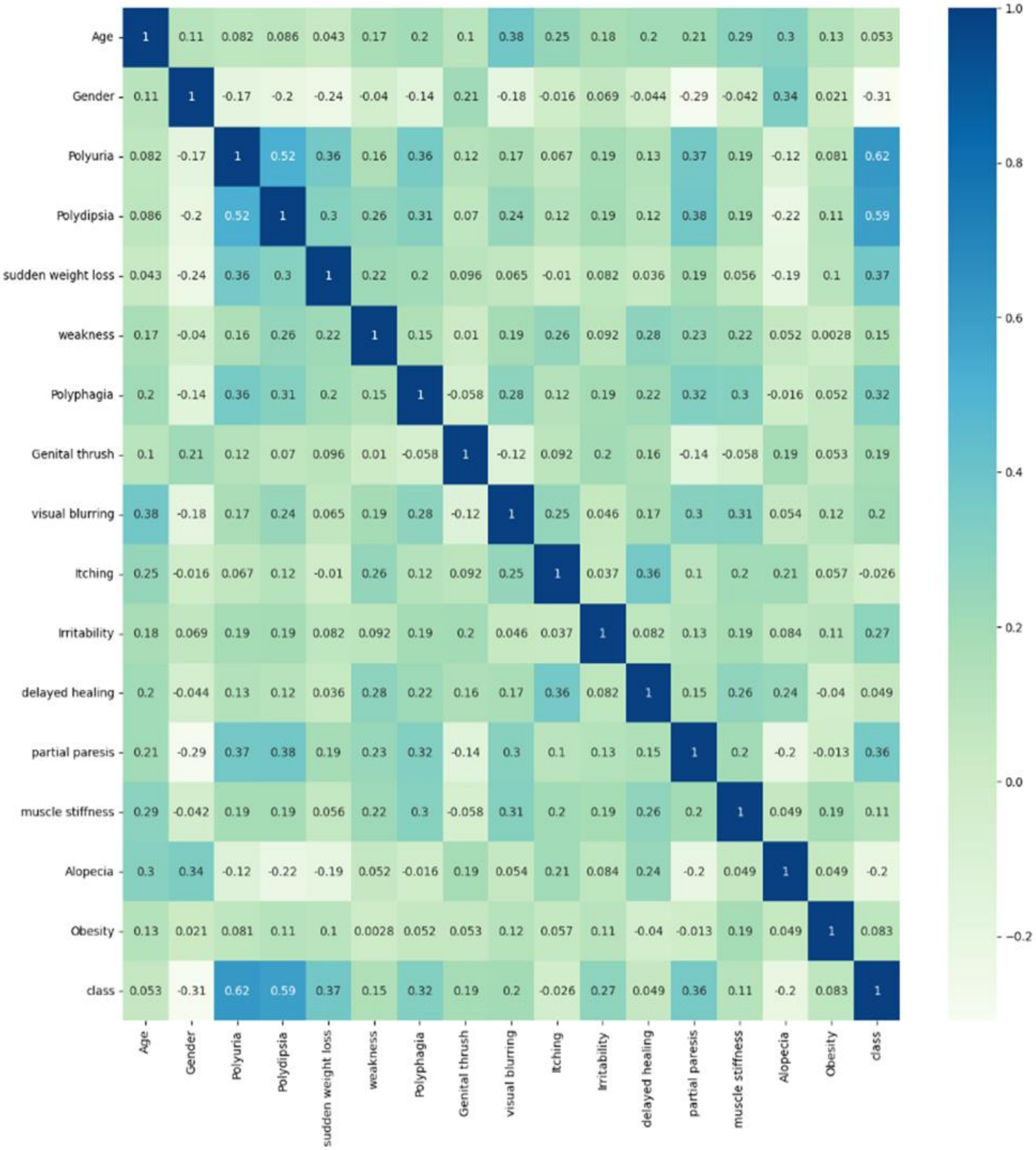


Figure 4: Correlation analysis of the attribute

2) Correlation Analysis

The data was mined to convert all values of "yes" to their corresponding binary value of 1 and "no" to 0 to create a correlation heatmap in Figure 4. Correlation maps visualize the relationship between two numerical variables to understand the extent to which these variables linearly relate to one another. Analyzing the correlation matrix, it becomes visible that the most substantial relationship with the outcome of the dataset is polyuria and polydipsia, with a factor of 0.62 and 0.59, respectively.

3) Splitting of Data

To train the model, a training and testing set of 70% and 30%, individually, is created by splitting the data, yielding 175 and 76 instances in the training and testing dataset, respectively.

C. Algorithms

Supervised machine learning methods are used to predict diabetes in its early stages. These underlying algorithms learn information from labeled data to train themselves. These trained algorithms are then given unlabeled test dataset to predict which category they fall into. Supervised machine learning deals with classification and regression problems. Since the outcome of our model is one of 0 or 1, this is a classification problem. The machine learning algorithms ' results are validated by utilizing a 5-fold cross-validation schema with training and testing sets organized as 175 and 76 samples for each fold. All data is divided into 5 folds for 5-fold cross-validation, and the model is trained on 4 folds on each turn while one fold is left for the testing of the model. This process is repeated 5 times. The machine learning techniques implemented on the dataset to achieve the desired outcome are stated below.

Insights: The study identifies SVM as the optimal algorithm with the highest F-score and recall for binary classification datasets, which highlights its practical utility for early diabetes risk prediction applications

1) Decision Tree

The decision tree classifier is a supervised and highly effective machine learning approach for classification. It entails making choices based on information from the past. Different parameters in a decision tree classifier form different tree nodes. The method selects a node at each stage by determining which property provides the most information gain. It splits data points into two related categories at a time, starting with the "tree trunk," continuing through the "branches," and ending with the "leaves" until the classes are more closely related to one another. This method is comparable to a flow chart. To enable organic classification without human intervention, categories inside categories are consequently constructed.

2) Random Forest

Using training data, many decision trees are initially created as part of the random forest method, which subsequently fits new data into one of the trees to form a "random forest." An ensemble of various decision trees makes up a random forest. The rationale behind random forest is to create decision trees by combining several sets of values from training sets, which lowers the likelihood of overfitting and misclassification by averaging the output of many decision trees. The random forest technique generates decision trees by taking several sets of values from the training set and combining them, which reduces the possibility of overfitting and inaccuracy, which tends to show by aggregating the performance of several decision trees.

3) Logistic Regression

To predict a binary outcome—that is, whether something happens or not—logistic regression is used. These results include Yes/No, 1/0, Pass/Fail, etc.

Analyzing independent variables yields the binary result, which belongs to one of two groups. Although the independent factors may be either category or quantitative, the dependent variable is often categorical.

4) Support Vector Machine

The best approach to categorizing the data is decided by a support vector machine based on the position of the border between the positive and negative classes. This boundary is known as the hyperplane because it maximizes the separation of data points from different categories. Support vector machines can be used to solve classification and regression issues, just like decision trees and random forests. Classification issues are dealt with using the SVC (support vector classifier).

5) Naïve Bayes

Naïve Bayes is based on Bayes' theorem [39]. Based on past knowledge of the circumstances surrounding an event, this theorem can describe the probability of that happening. Although a class's features may be interdependent, this classifier presumes that a specific element in the class is not directly related to any other feature [40]. Naïve Bayes has different algorithms such as Gaussian Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and others.

6) K-Nearest Neighbor

One of the earliest and most straightforward classification techniques is the K-nearest neighbor (KNN) algorithm. [41]. A detailed but efficient machine learning approach is the K-nearest neighbor. KNN algorithm surmises the similar features of the new data and the data already fed to the machine. It works on memory and makes the classification of new data points on the similarity of previous cases. This means that using the KNN method, fresh data can be quickly sorted into the most suitable category when it first arises [42].

7) XG Boost

Extreme gradient boosting begins by creating weak models and ends with a robust model. It is a framework of tree-based machine learning. More nodes are added to decision trees in parallel while considering the gradient of the loss function. Extreme gradient boosting is a tree-based machine learning framework that starts with weak models and finishes with a strong model. More nodes are added to decision trees in parallel while considering the gradient of the loss function. Each tree's outcome is considered when sorting an instance, and the findings with the highest count are returned as the model's output.

8) AdaBoost

Adaptive Boosting, AdaBoost, is an ensemble method of boosting techniques. It can work with many other learning algorithms, such as decision trees in conjunction, to increase accuracy performance. It is usually used in binary classification but can also be generalized with multiple classes. The approach is characterized as adaptive boosting since each instance is given a fresh set of weights, with cases that were incorrectly classified receiving higher weights. It decreases the loss function and can solve differentiable loss function problems. AdaBoost is used for the classification as well as for regression.

PERFORMANCE EVALUATION METRICS

The most used diagnostic tools for performance evaluation of classifiers are accuracy, recall, precision, and F-score. Accuracy is the metric used to determine which model most effectively identifies connections and patterns between variables in a dataset based on the input or training data. However, accuracy does not give detailed information about model performance. Another indicator of a model's performance is the ratio of true positive predictions made by the model over all of the positive classified data, or precision. A model with high precision classifies diabetic patients accordingly without marking healthy people as diabetic. For an ideal classifier, precision is 1. From eq (2), it can be seen that precision is 1 only when the number of false positives is zero, i.e., when the numerator and denominator values are equal. As FP increases, there is an imbalance in the numerator and denominator values, and the precision decreases. The proportion of correctly classified positive samples to all positive samples is used to calculate a model's recall or sensitivity. Precision can be considered a quality metric, whereas recall, a quantity metric.

Similar to precision, an ideal recall is 1. From eq (3) it is visible that recall becomes 1 only when the number of false negatives is zero. The higher the FN, the lower the recall. The F-score, or the F1 score or F-measure, considers FN and FP and is a harmonic mean of the system's precision and recall. The F score reaches 1 only when both precision and recall are 1, as evident from eq(5).

Model performance can also be evaluated by a confusion matrix, an error, or a contingency matrix. The confusion matrix provides additional details regarding a classification model's performance, including whether classes are correctly or incorrectly predicted and mistakes the model produces. An n x n matrix where n denotes the number of target classes. For a binary classification system, there are four parameters, namely;

- The correctly predicted values are Positive (TP) and True Negative (TN).
- The incorrectly predicted values are False Positive (FN) and False Negative (FN).

Predicted Values	Actual Values Positive	Actual Values Negative
Negative	TP	FP
Positive	FN	TN

Figure 5 Parameters of the confusion matrix

The TPR and FPR are shown against various threshold values to create the ROC (Receiver Operating Characteristics) curve. The area under the ROC curve (AUC) governs the predictability of a classifier [43]. This shows how capable the model is in differentiating between classes. The ability of the classifier to distinguish between diabetic patients and healthy individuals improves with increasing AUC in diabetes classification. The ROC curve is plotted against TPR (True positive rate) and FPR (False positive rate). The FPR gives the number of incorrect predicted values in the positive class. For better performance, TPR should be high, whereas FPR should be low. Following are all the formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = True\ positive\ rate = \frac{TP}{TP + FN} \quad (3)$$

$$False\ positive\ rate = \frac{FP}{FP + TN} \quad (4)$$

$$F1\ score = 2 * \frac{P * R}{P + R} \quad (5)$$

RESULTS AND DISCUSSIONS

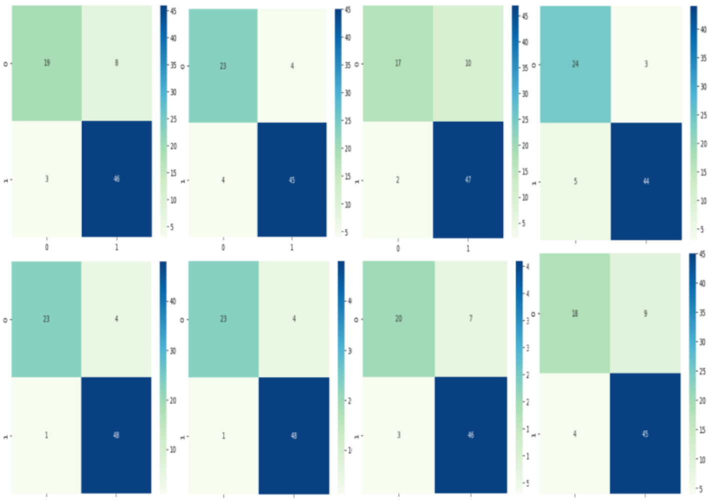


Figure 6: From left to right: Decision Tree, Naive Bayes, Logistic Regression, KNN, SVM, Random Forest, XG Boost, AdaBoost

Different results were obtained as the working criteria of all the aforementioned algorithms differ. The models' accuracy was projected using their confusion matrices, as shown in Figure 6. The results evaluated based on their accuracy, recall, precision, and f-score are shown in Table III. The training and testing sets were structured as 175 and 76 samples for each fold in a 5-fold cross-validation schema used to validate the results. According to the results, both SVM and Random Forest have the highest accuracy of 93.4%. The classifier with the lowest accuracy is AdaBoost, with an accuracy of 82.8%. A comparative analysis of all of the classifiers is given in Figure 7. In terms of precision, Random Forest had the highest precision of 94%, which means that the probability of Random Forest classifying a healthy patient as diabetic is the lowest of all other models. Although all the classifiers performed well during recall, the highest recall score is of SVM. A fundamental metric, the F-score, balances recall and precision, and the model with the highest f1 score is SVM.

Table III: The effectiveness of the machine learning classifiers used for predicting the likelihood of developing early-stage diabetes

Model	Accuracy (%)	Recall (%)	Precision (%)	F score (%)
Decision Tree	85.5	93.8	85.1	89.3
Naïve Bayes	89.4	91.8	91.8	91.8
Logistic Regression	84.2	95.9	82.4	88.6
KNN	89.4	89.7	93.6	91.6
SVM	93.4	97.9	92.3	95
Random Forest	93.4	95.9	94	94.9
XG Boost	86.8	93.8	86.7	90.1
AdaBoost	82.8	91.8	83.3	87.3

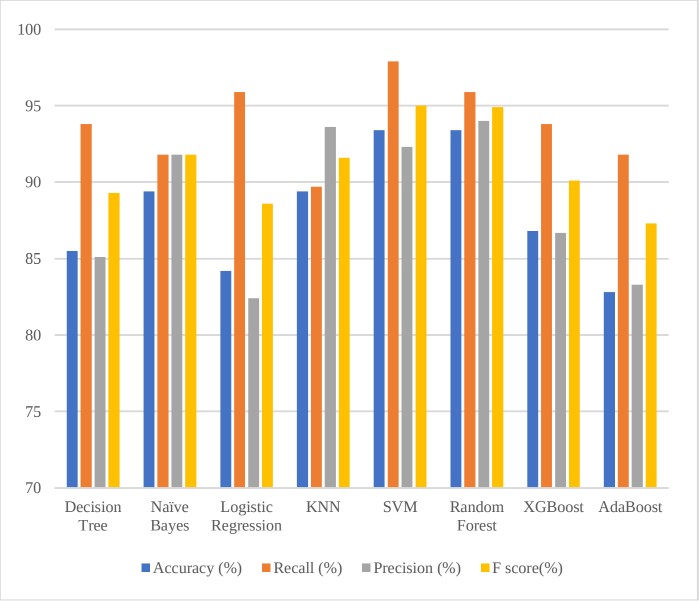
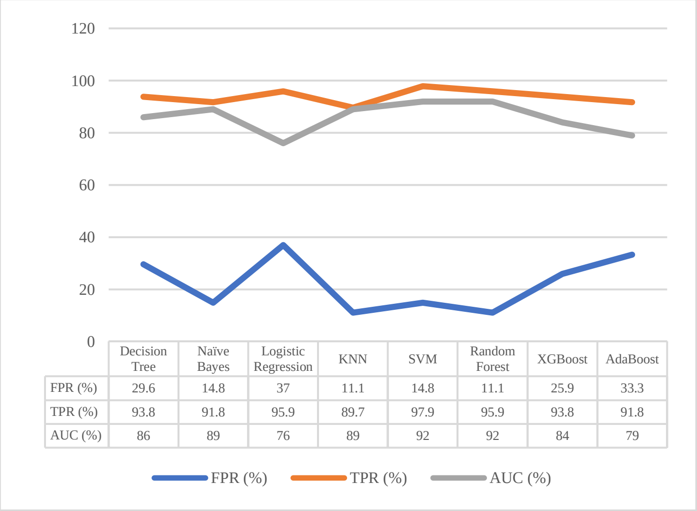


Figure 7: Comparative evaluation of the machine learning methods used for predicting the probability of developing diabetes



	Decision Tree	Naïve Bayes	Logistic Regression	KNN	SVM	Random Forest	XGBoost	AdaBoost
FPR (%)	29.6	14.8	37	11.1	14.8	11.1	25.9	33.3
TPR (%)	93.8	91.8	95.9	89.7	97.9	95.9	93.8	91.8
AUC (%)	86	89	76	89	92	92	84	79

Figure 8: Comparative evaluation of FPR, TPR, and AUC of the machine learning methods used for predicting the probability of developing diabetes

A comparative evaluation of FPR, TPR, and AUC is given in Figure 8. A model with high TPR and low FPR is considered well-performing. Of all the classifiers, SVM had the highest TPR score with an FPR score of 14.8%, while KNN had the lowest FPR score with a TPR score of 89.7%. SVM and Random Forest both have the highest AUC of 92%.

Comprehensive Evaluation: The inclusion of comparative literature highlights the paper's contributions, showing its superiority over existing algorithm

CONCLUSION

Millions worldwide suffer from diabetes, a chronic and fatal health condition. Consequently, it is essential to identify diabetes early on. This study used eight machine learning algorithms on a dataset of 520 instances and 17 characteristics to produce eight models for predicting the probability of developing early-stage diabetes. All except one feature are continuous. All categorical variables are denoted by 1 (for positive) and 0 (for negative). The performance evaluation metrics used were accuracy, recall, precision, f1, and AUC score. Both RF and SVM had the highest accuracy and AUC scores. However, SVM had a higher TPR score than RF and a higher recall and F score. Since evaluated metrics for SVM are more elevated, further work on creating an early-stage risk prediction application may be done using the developed Support Vector Machine model.

ABBREVIATIONS

ACC	Accuracy
AUC	Area Under the Curve
ANN	Artificial Neural Networks
CM	Confusion Matrix
CNN	Convolutional Neural Network
DT	Decision Tree
EMC	Electronic Medical Claim
FPR	False Positive Rate
FKNN	Fuzzy K Nearest Neighbour
GBM	Gradient Boosting Machine
GRNN	Generalized Regression Neural Network
J48	Java 48
JRIP	Ripper
KNN	K-Nearest Neighbors
LS-SVM	Least-Squares Support-Vector Machines
LR	Logistic Regression
MCC	Matthew's Coefficient Correlation
MLP	Multilayer Perceptron
MLPNN	Multi-Layer Perceptron Neural Network Model
MRI	Magnetic Resonance Imaging
NB	Navies Bayes
NBTree	Naïve Bayes Tree
NPV	Negative Predictive Value
NN	Neural Network
PIDD	Pima Indian Diabetes Dataset
PPV	Positive Predictive Value
PRC	Precision–Recall Curve
RF	Random Forest Classifier.
ROC	Receiver Operating Characteristic Curve
SVM	Support Vector Machine
SEER	Surveillance, Epidemiology, and End Results
TPR	True Positive Rate
UCI	University of California, Irvine

REFERENCES

[1] A. Sapra and P. Bhandari, "Diabetes Mellitus," *StatPearls*, Jun. 2022, Accessed: Nov. 27, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK551501/

[2] Diabetes Mellitus - StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK551501/ (accessed Nov. 27, 2022).

[3] Top 10 Most Common Health Issues in Pakistan - Healthwire. https://healthwire.pk/healthcare/common-health-issues-in-pakistan/ (accessed Nov. 27, 2022).

[4] The top 10 causes of death." https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed Nov. 27, 2022).

[5] Federation, I. D. (2013). Five questions on the IDF Diabetes Atlas. Diabetes research and clinical practice, 102(2), 147-148.

[6] Facts & figures." https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html (accessed Nov. 27, 2022).

[7] Iqbal, M. W., Naqvi, M. R., Khan, M. A., Khan, F., & Whangbo, T. (2022). Mobile Devices Interface Adaptivity Using Ontologies. Computers, Materials & Continua, 71(3).

[8] Khan, M. A., Abbas, S., Raza, A., Khan, F., & Whangbo, T. (2022). Emotion Based Signal Enhancement Through Multisensory Integration Using Machine Learning. Computers, Materials & Continua, 71(3).

[9] Ayvaz, U., Gürüler, H., Khan, F., Ahmed, N., Whangbo, T., & Bobomirzaevich, A. A. (2022). Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning. Computers, Materials & Continua, 71(3).

[10] Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes, 15(3), 435-443.

[11] Tariq, H., Rashid, M., Javed, A., Zafar, E., Alotaibi, S. S., & Zia, M. Y. I. (2021). Performance analysis of deep-neural-network-based automatic diagnosis of diabetic retinopathy. Sensors, 22(1), 205.

[12] Kumar, D., Jain, N., Khurana, A., Mittal, S., Satapathy, S. C., Senkerik, R., & Hemanth, J. D. (2020). Automatic detection of white blood cancer from bone marrow microscopic images using convolutional neural networks. IEEE Access, 8, 142521-142531.

[13] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. International Journal of Cognitive Computing in Engineering, 2, 40-46.

[14] Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). [Retracted] A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey. Journal of Healthcare Engineering, 2022(1), 8100697.

[15] Chatrati, S. P., Hossain, G., Goyal, A., Bhan, A., Bhattacharya, S., Gaurav, D., & Tiwari, S. M. (2022). Smart home health monitoring system for predicting type 2 diabetes and hypertension. Journal of King Saud University-Computer and Information Sciences, 34(3), 862-870.

[16] Ani, R., Sasi, G., Sankar, U. R., & Deepa, O. S. (2016, September). Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1287-1292). IEEE.

[17] Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, 8, 1-14.

[18] Lynch, M. R., Tran, M. T., & Parikh, S. M. (2018). PGC1α in the kidney. American Journal of Physiology-Renal Physiology, 314(1), F1-F8.

[19] Chen, C. Y., Su, C. H., Chung, I. F., & Pal, N. R. (2012, June). Prediction of mammalian microRNA binding sites using random forests. In 2012 International Conference on System Science and Engineering (ICSSE) (pp. 91-95). IEEE.

[20] Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.

[21] Eskidere, Ö., Ertaş, F., & Hanilçi, C. (2012). A comparison of regression methods for remote tracking of Parkinson’s disease progression. Expert Systems with Applications, 39(5), 5523-5528.

[22] Yadav, D. C., & Pal, S. (2021). An experimental study of diversity of diabetes disease features by bagging and boosting ensemble method with rule based machine learning classifier algorithms. SN Computer Science, 2(1), 50.

[23] Chen, H. L., Huang, C. C., Yu, X. G., Xu, X., Sun, X., Wang, G., & Wang, S. J. (2013). An efficient diagnosis system for detection of Parkinson’s disease using fuzzy k-nearest neighbor approach. Expert systems with applications, 40(1), 263-271.

[24] Goyal, P., & Jain, S. (2022, March). Prediction of type-2 diabetes using classification and ensemble method approach. In 2022 International Mobile and Embedded Technology Conference (MECON) (pp. 658-665). IEEE.

[25] Prakash, A. (2021). An ensemble technique for early prediction of type 2 diabetes mellitus–a normalization approach. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), 2136-2143.

[26] Doğru, A., Buyrukoğlu, S., & Arı, M. (2023). A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing, 61(3), 785-797.

[27] Behroozi, M., & Sami, A. (2016). A Multiple‐Classifier Framework for Parkinson’s Disease Detection Based on Various Vocal Tests. International journal of telemedicine and applications, 2016(1), 6837498.

[28] Ergün, Ö. N. (2021). Early stage diabetes prediction using machine learning methods. Avrupa Bilim ve Teknoloji Dergisi, (29), 52-57.

[29] Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A novel approach for feature selection and classification of diabetes mellitus: machine learning methods. Computational Intelligence and Neuroscience, 2022(1), 3820360.

[30] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716.

[31] Reddy, D. J., Mounika, B., Sindhu, S., Reddy, T. P., Reddy, N. S., Sri, G. J., ... & Kora, P. (2020).

Predictive machine learning model for early detection and analysis of diabetes.

[32] Hussain, L., Ahmed, A., Saeed, S., Rathore, S., Awan, I. A., Shah, S. A., ... & Awan, A. A. (2018). Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. Cancer Biomarkers, 21(2), 393-413.

[33] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 77(5), 5198-5219.

[34] Raghavendran, C. V., Naga Satish, G., Kumar Kurumeti, N. S. L., & Basha, S. M. (2022). An Analysis on Classification Models to Predict Possibility for Type 2 Diabetes of a Patient. In Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2021 (pp. 181-196). Singapore: Springer Nature Singapore.

[35] Laila, U. E., Mahboob, K., Khan, A. W., Khan, F., & Taekeun, W. (2022). An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. Sensors, 22(14), 5247.

[36] Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R., & Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial intelligence in medicine, 20(1), 59-75.

[37] Hung, C. Y., Chen, W. C., Lai, P. T., Lin, C. H., & Lee, C. C. (2017, July). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3110-3113). IEEE.

[38] Dutta, I. (2020). Early stage diabetes risk prediction dataset. Kaggle.

[39] Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. Journal of the Royal Statistical Society. Series B (Methodological), 102-107.

[40] “[PDF] An empirical study of the naive Bayes classifier | Semantic Scholar.” https://www.semanticscholar.org/paper/An-empirical-study-of-the-naive-Bayes-classifier-Watson/2825733f97124013e8841b3f8a0f5bd4ee4af88a (accessed Nov. 27, 2022).

[41] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[42] Point, J. T. (2023). K-nearest neighbor (knn) algorithm for machine learning.

[43] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16.