# Diabetes Prediction Using Machine Learning Methods

**Nahla Aljojo[1], Ghadah Alharbi[2], Amani Alraddadi[3], Noura Alshareef[4], Noura almeheyawil[5]**

*"College of Computer Science and Engineering, Information system and Technology Department, University of Jeddah, Jeddah, Saudi Arabia"*

**ABSTRACT** Machine learning algorithms anticipate medical data at early stages of safe human life. Large amounts of medical data can be found in various data sources and used in real-world apps. Today, ML answers questions, particularly in predicting disease data. In addition, a large amount of data generated by healthcare biotechnology can be used very efficiently to promote healthy living . Accordingly, early detection of diabetes can be achieved with the help of data mining prediction technology, by which diabetes data analysis can be provided. In this paper, disease probability can be predicted by providing a comparative study in which the most popular machine-learning techniques are discussed, whish is Random Forests (RF) and Support Vector Machines (SVM). In performance metrics, The random forest model achieved an accuracy of 79% and was outperformed by the support vector machine with an accuracy of 84% . Furthermore, performance metrics include F-measures that can be provided by the confusion matrix, precision, accuracy, and recall.

INDEX TERMS *Diabetes; Random Forests (RF); SVM; Accuracy; Machine Learning.*

## I INTRODUCTION

The “World Health Organization” (WHO) indicated that one of the rapidly growing diseases is diabetes which is considered to be a chronic disease. In addition, irregular levels of blood sugar are caused due to diabetes by which many parts of the body are damaged, including blood vessels, the nerves, eyes, and heart. Moreover, the source of energy for the cells of the body is glucose which can be moved from the blood into the cells with the help of a hormone that is called “insulin” which is produced by the pancreas. If enough “insulin” can't be produced by the pancreas, then diabetes occurs. In addition, diabetes can occur because insulin can't be used effectively by the body. Consequently, serious health problems can be caused because the cells cannot utilize the excess sugar in the blood [1].

Diabetes has two types. To clarify, insufficient production of insulin by the pancreas is the reason for type1 diabetes to occur. When it comes to type 2 diabetes, it's proven that the human life can be seriously affected by this disease which is caused by insulin resistance, and insufficient production of insulin by the pancreas [1].

Machine learning encompasses the creation of algorithms and methods that equip computers with the capability to gain knowledge and become intelligent based on previous experiences. This field of research is part of "artificial intelligence" (AI) and has a strong connection to statistics. The learning process helps the system understand and recognize input data, allowing it to make decisions and predictions based on that information.[2]

The learning steps starts by collecting data from different sources. Then, the data is cleaned and organized, and unnecessary information is removed. This is called preprocessing. Since there is a lot of data, it can be hard for the computer to make decisions, so special methods (using logic, probability, statistics, etc.) are used to help it understand the data. The computer is then tested to see how well it is doing and to see if there is anything that can be improved. Finally, the computer is made better by adding new information or rules. Machine learning is used to categorize things, guess what might happen next, and find patterns. [2].

In this research, we employed supervised machine learning techniques through SAS to analyze the Pima Indian diabetes dataset, which is publicly accessible on "Kaggle [3]. The purpose of our model was to uncover

patterns for knowledge discovery in diabetes. The study will examine the performance of Random Forests (RF) and Support Vector Machines (SVM). Our findings indicate that SVM produced the highest accuracy rate of 84%, while Random Forest (RF) had a rate of 79%. While multiple machine learning techniques can be used for predictions, determining the best approach is not straightforward. Furthermore, by applying data mining, we can provide a tool for early diabetes diagnosis, which has the potential to save lives in the future.

## II RELATED WORK

Machine learning algorithms are widely recognized in the medical field for disease prediction. Researchers often utilize ML methods for the most accurate prediction of diabetes. This research includes a compilation of 10 studies related to machine learning, as summarized in (Table 1). Half of them are using support vector machine algorithm and random forest, in addition to studies in deep learning.

In paper [4], the authors compared the performance of the ontology classifier with various machine learning algorithms using cross-validation and percentage split evaluation criteria. The ontology classifier showed the highest accuracy of 77.5%, followed by SVM (77.3%) and Logistic Regression (77.2%).

In [5], the authors used random forest, multilayer perceptron, and logistic regression classifiers and long short-term memory, moving averages, and linear regression predictive models for diabetes classification and prediction using the "PIMA Indian Diabetes dataset". The multilayer perceptron classifier had the highest accuracy of 86.08%, while the long short-term memory model achieved the highest accuracy of 87.26% in predictive analysis.

In [6], the authors examined the prediction accuracy of the Support Vector Machine (SVM) algorithm with different kernels for diabetes prediction in Indian patients. The RBF kernel was found to have the highest prediction accuracy of 82%.

In [7], the authors compared five supervised learning techniques, including K-Nearest Neighbors (K-NN), Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting, for diabetes prediction using the "Pima-Indian dataset". The highest accuracy of 0.81 was achieved by Gradient Boosting, while K-NN had the lowest accuracy of 0.64.

The author in [8], implemented a deep neural network for diabetes prediction and found an accuracy of 98.35% for five-fold cross-validation, which was higher than other methods.

In reference [9], the authors propose a diabetes prediction system and compare two traditional ML methods (Support Vector Machine (SVM) and Random Forest (RF)) with a deep learning method using a Convolutional Neural Network (CNN). The results show that Random Forest had the best accuracy at 83.67%, followed by Deep Learning at 76.81% and Support Vector Machine at 65.38%.

In [10], the authors used various tools to study important factors and perform data analysis on diabetes. They discovered a strong connection between diabetes and factors such as body mass index (BMI) and blood sugar levels. They used (Artificial Neural Network (ANN), Random Forest (RF), and K-means clustering) to predict diabetes and found that ANN had the highest accuracy of 75.7%.

In [11], Deepti and Dilip evaluated the accuracy of (Decision Tree, SVM, and Naive Bayes classifiers) for diabetes detection using the Pima Indian dataset and 10-fold cross-validation. The best accuracy rate, at 76.30%, was achieved by using the Naive Bayes classifier.

In [12], the authors applied various classification methodologies, including (Naive Bayes, Neural Network, AdaBoost, kNN, Random Forest, and SVM), to a public diabetes dataset. The Neural Network method was found to be the most effective with 98.1% accuracy in predicting early stages of diabetes.

In [13], Mercaldo et al. method to distinguish between diabetic and non-diabetic patients using machine learning. they train the model with 6 different classification algorithms and achieve a precision of 0.757 by HoeffdingTree algorithm after selecting the best features.

**Table (1): Summary of the Machine Learning Algorithm related works.**

| Ref. | years | goal | algorithm | Accuracy |
|---|---|---|---|---|
| [4] | 2022 | A comparative review of the top machine learning methods and ontology-based machine learning classification. | ontology classifier<br>SVM<br>Logistic Regression | ontology :77.5%,<br>SVM: 77.3%<br>Logistic: 77.2% |
| [5] | 2021 | Using various ML techniques to predict the likely presence of diabetes at an early stage -- especially in women -- | MLP algorithm | The accuracy:86% |
| [6] | 2020 | applied SVM. The prediction accuracy depends on the selection its model. | SVM ML algorithm | RBF model (82%) |
| [7] | 2019 | How accurate the method we're going to use will allow us to find out which of the 5 methods is the most accurate and which is the least accurate. | five supervised learning techniques | High Accuracy is<br>Gradient Boosting: 81% |
| [8] | 2019 | Use deep Learning Approach with five-fold cross-validation. | Deep Learning<br>Neural Network | The accuracy 98.35% |
| [9] | 2019 | Efficiency Analyzing of Traditional ML and Deep Learning Methods for Diabetes Prediction | Random Forest<br>Deep Learning<br>SVM | Random Forest: 83.67%<br>Deep Learning:76.81%<br>SVM: 65.38% |
| [10] | 2019 | Determine significant attribute selection, and use clustering, prediction, and association rule mining for diabetes. | (ANN), (RF)<br>K-means | ANN: 75.7%<br>Rf: 74.7%<br>K-means 73.6% |
| [11] | 2018 | Design a model that most accurately predicts a patient's likelihood of developing diabetes. | SVM - Naive Bayes - Decision Tree | High Accuracy is<br>naive bayes 76.30% |
| [12] | 2018 | Comparison of data mining methods for the early detection of diabetes | SVM,KNN,ANN,<br>Naïve Bayes,<br>Logistic regression<br>and Decision Tree. | High Accuracy is<br>ANN: 98.4% |
| [13] | 2017 | Classification and Diagnosis of Diabetes Mellitus Patients using ML | HoeffdingTree algorithm | The accuracy :75% |

## III PROPOSED METHODOLOGY

### *Dataset*

The dataset used in this study was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The goal is to use diagnostic measures to predict whether a patient has diabetes. (publicly available at [3]). There are several limitations to selecting these instances from larger databases. Specifically, all the patients here were women over the age of 21 who were of Pima Indian descent. There are 768 cases and 9 numeric attributes associated with diabetes. Of the available data, 60% of the data is used as training data, the remaining 10% is used for testing, and 10% is used for validation. Table 2 gives a detailed description of all attributes.[14].

**Table 2: Dataset description table**

| No. Attributes | Attributes | Attribute value and Descriptions |
|---|---|---|
| 1 | Pregnancies | number of times pregnant |
| 2 | Glucose | Numerical Values, Plasma glucose concentration a 2 hour in an oral glucose tolerance test |
| 3 | Blood Pressure | Numerical Values, Diastolic blood pressure (mm Hg) |
| 4 | Skin Thickness | Numerical Values, Triceps skin fold thickness |
| 5 | Insulin | Numerical values, (mu / U/ml) |
| 6 | Body Mass Index "BMI" | Numerical values, in (weight in Kg / (height in m) ^ 2 |
| 7 | Diabetes Pedigree Function | Numerical values, Diabetes Pedigree Function |
| 8 | Age | Numerical values, (years) |
| 9 | Outcome | Yes = 1<br>No = 0 |

### *A visualization*

Data visualization is the graphical representation of information and data.Finding problem areas, improving decisions based on data-driven insights, and successfully communicating findings can all be helped by data visualization.

### *B Histogram*

A histogram is a graphical representation of data that uses bars of different heights to show how many times a value occurs within a set of data. It is used to display the distribution of numerical data. The x-axis typically represents the categories or values being measured, and the y-axis represents the frequency or count of those values. seen in ( Figure 1) Also, ( Figure 2) .

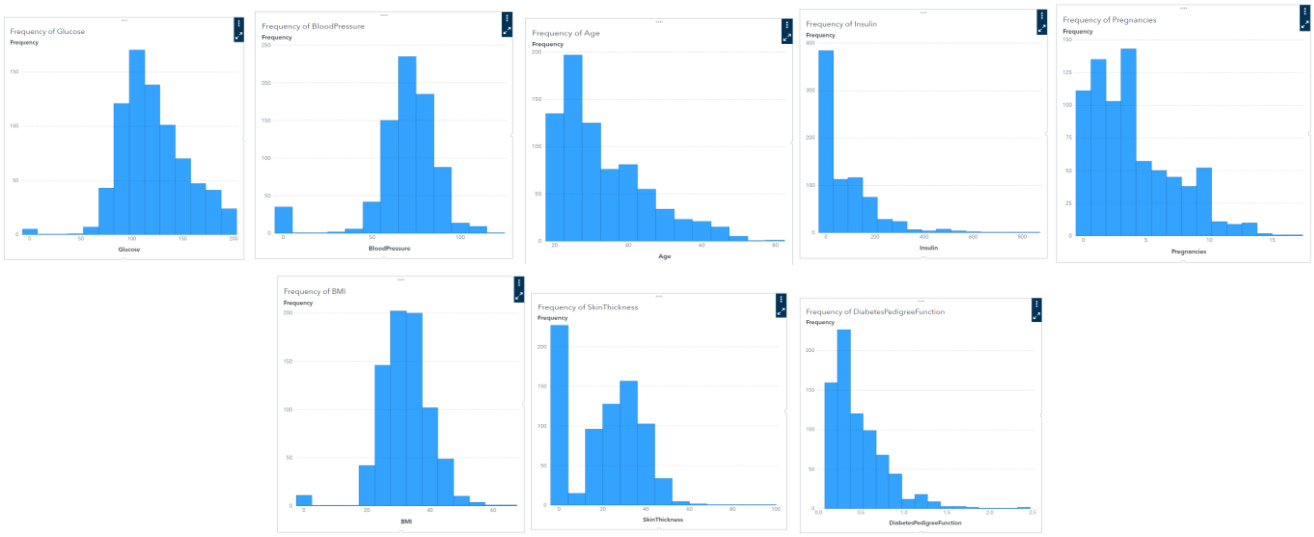


**FIGURE 1: Histogram of features**

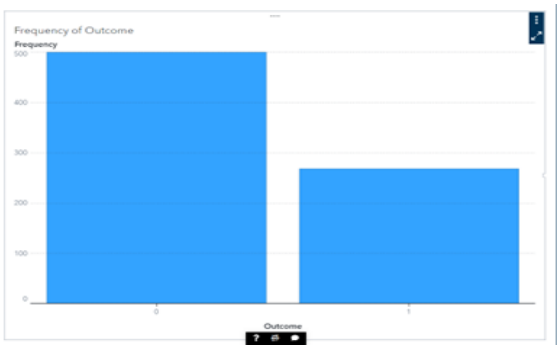


**FIGURE 2: Bar Chart For Outcome Class**

### *D Boxplot*

A box plot is a visual representation of the distribution of numerical data that allows for comparisons between variables or between levels of a category variable. Except for points that are identified as "outliers" using a method that is a function of the inter-quartile range, the box displays the dataset's quartiles, and the whiskers expand to display the remainder of the distribution. seen in (Figure 3).

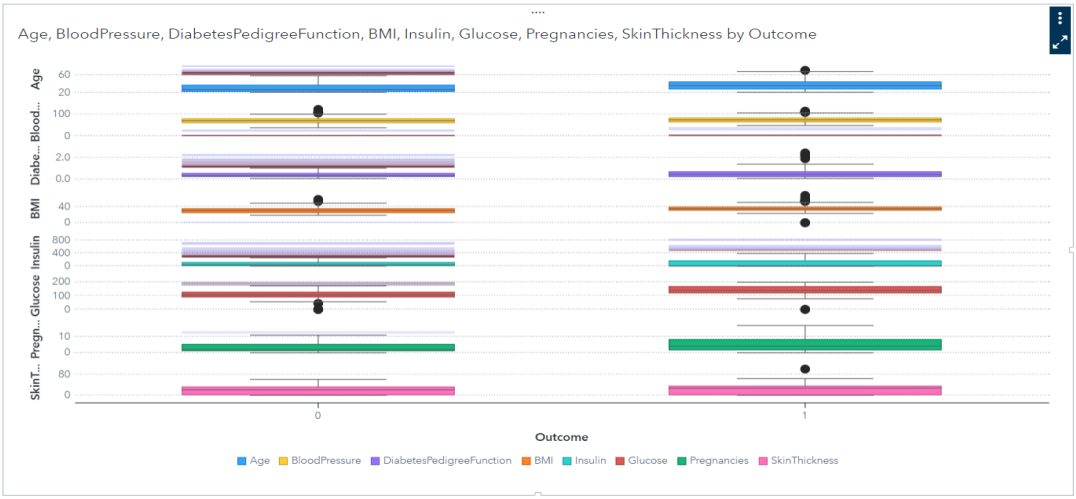


**FIGURE 3: Description of data preprocessing steps and justifications**

### *Data preprocessing*

Real-world data may have noisy, inconsistent, or missing values. No quality results might be found if the data quality is poor. To produce high-quality results, the data must be preprocessed. The data is preprocessed using cleaning, integration, transformation, reduction, and discretization.

### *Data Cleaning*

is the process of preparing data for analysis by removing or modifying inaccurate, incomplete, or irrelevant parts, data cleaning is an essential part of any data analysis project as it ensures that the data is accurate and reliable.In our data there are no missing values.

## IV MODEL BUILDING ALGORITHMS

Model building is the process of creating a mathematical representation of a real-world system. It involves identifying the relationships between different variables and using those relationships to create a model that can be used to make predictions or decisions. The process typically involves collecting data, analyzing it, and then constructing a model based on the results. Once the model is built, it can be tested and refined until it accurately reflects the real-world system it is intended to represent.seen in (Figure 4).

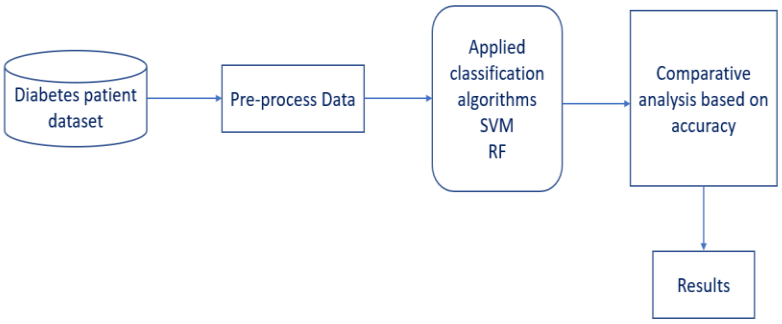


**FIGURE4: Model Building**

### *A Support Vector Machine (SVM):*

One of the common supervised machine learning models for classification is the SVM. The SVM aims to determine the optimal separating hyperplane with the highest border between the two classes given a training sample of two classes.

Performance of an SVM algorithm using a confusion matrix is as (Figure 5)

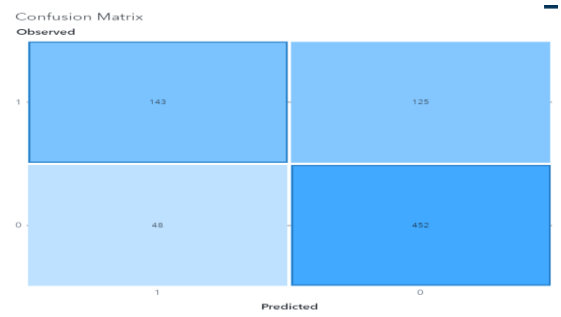


**FIGURE 5: confusion matrix in SVM**

The Support Vector Machine assigns a value to each feature based on its relevance. Each attribute is assigned a number between 0 and 1, with 0 denoting "not used at all" and 1 denoting "exactly predicts the target.".as seen in (Figure 6) “Glucose” is unquestionably the most crucial component.

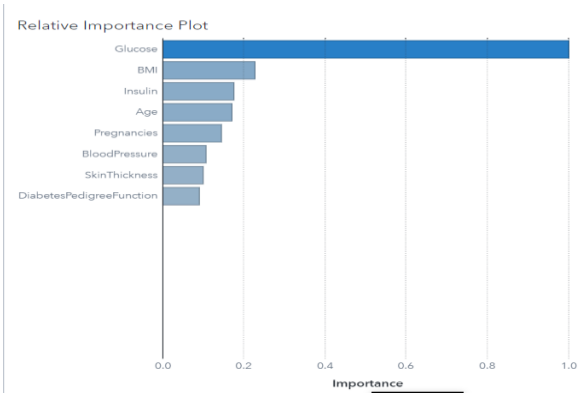


**FIGURE 6: Significance of the features.**

### *B Random Forest (RF)*

Random Forest is an ensemble machine learning algorithm that uses a collection of decision trees to make predictions. It works by randomly selecting a subset of features from the data and then building multiple decision trees based on those features. The final prediction is made by averaging the predictions from each tree. Performance of an RF algorithm using a confusion matrix seen in (Figure 7)

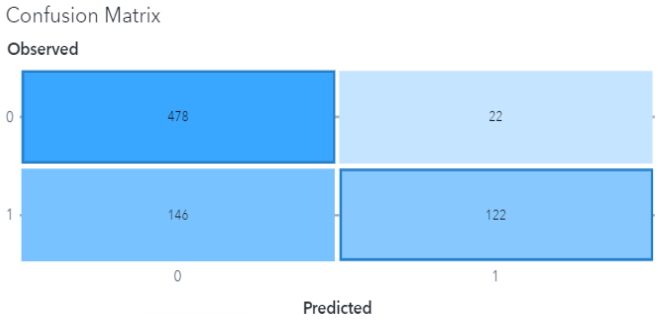


**FIGURE 7: confusion matrix for Random Forest.**

### *C Feature importance in Random Forest:*

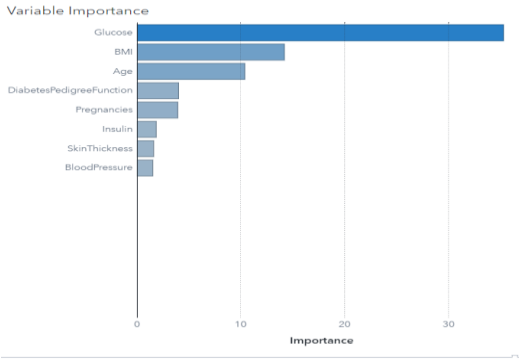


**FIGURE 8: Feature "Glucose" is by far the most important feature.**

### *D correlation matrix*

The correlation matrix as seen in (Figure 9) makes it easier to see which predictive analytics techniques are more effective in identifying the best model. As can be shown, the admission rate has a positive correlation with all of the parameters. The value of "glucose" is the feature that has the biggest effect on admission rates, with a significant value of R=0.46, while the value of "blood pressure" is the feature with the smallest

significant value of R=0.16.

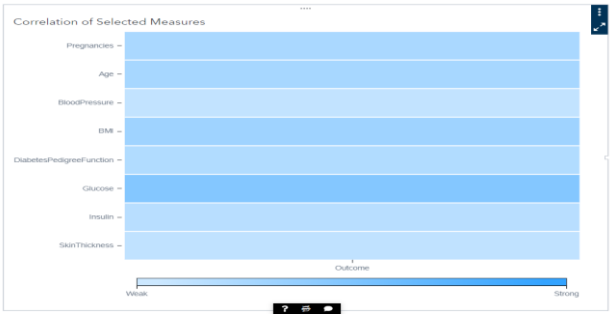


**FIGURE 9: Feature correlation matrix**

## V MODEL BUILDING

The champion model for this project is SVM. The model was chosen based on the KS (Youden) for the Test partition (0.61). 84.42% of the Test partition was correctly classified using the SVM model. The five most important factors are Glucose. Not high (outlier, kurtosis, skewness) - five bin decision tree binning, Glucose: Not high (outlier, kurtosis, skewness) - log + impute(median), BMI: Not high (outlier, kurtosis, skewness) – power (2) + impute(median), BMI: Not high (outlier, kurtosis, skewness) - five bin decision tree binning, and Insulin: Low missing rate - median imputation.

### *A Accuracy Measures:*

This research work makes use of algorithms. Tenfold internal cross-validation is used throughout experiments. This study is categorized using accuracy, F1 score, recall, precision, and ROC (Receiver Operating Curve) measurements. Below are the accuracy metrics defined in Table 2.

**TABLE 2. Accuracy Measures**

| Measures | Definitions | Formula |
|---|---|---|
| **Accuracy (A)** | measure of how well a model correctly predicts the outcome of a given data set. | A=(TP+TN) / (Total |
| **Precision (P)** | measure of how accurately a model predicts the outcome of a given data set. | P = TP / (TP+ FP) |
| **Recall (R)** | **MEASURE OF HOW MANY RELEVANT RESULTS ARE RETURNED BY A MODEL.** | R =TP / (TP+FN) |
| **F1 score** | F1 is a measure of accuracy that combines Precision and Recall. | F=2*(P*R) / (P+R) |
| **ROC** | ROC (Receiver Operating Curve) curve is used to evaluate the performance of classification models. | |

Table 3 lists the corresponding classifiers' performance in terms of Accuracy, Precision, F1 score, Recall, and ROC values and defines classifier performance in terms of classified instances.
Where TP, TN, FP, FN, and FP define True Positive, True Negative, True Negative, and False Positive, respectively.

**TABLE 3. Comparative Performance of Classification Algorithms on Various Measures.**

| Classification Algorithms | Precision | Recall | F1 score | Accuracy % | ROC |
|---|---|---|---|---|---|
| SVM | 53.35 | 74.86 | 72.73 | 84.42 | 0.5726 |
| RF | 95.6 | 76.60 | 63.64 | 79.22 | 0.4585 |

## VI RESULTS

Table 3 represents different performance values of all classification algorithms calculated on various measures. From Table-3 it is analyzed that SVM showing the maximum accuracy. So the SVM machine learning classifier can predict Compared to other classifiers, this one accurately predicts the likelihood of diabetes. Figure 11 represents the ROC area of all classification algorithms.

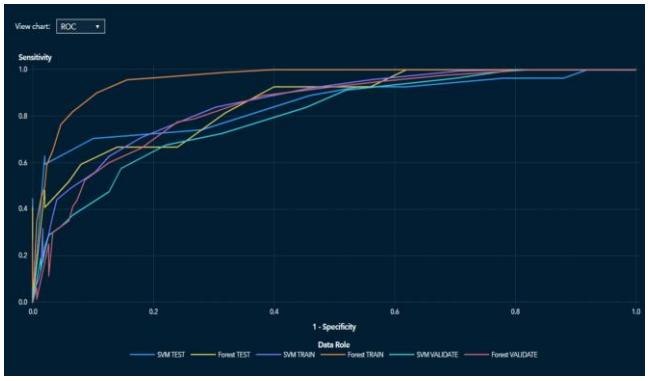


**FIGURE 11: ROC Area of all Classification Algorithms**

## VII CONCLUSION AND FUTURE WORK

The goal of this study was to predict diabetes using “machine learning techniques”, including “Random Forest (RF) and Support Vector Machine (SVM)” . The results showed that SVM was the most accurate with 84% and Random Forest was 79%. To improve the accuracy of predictions, the study suggests incorporating deep learning techniques because previous research has shown that deep learning can produce accurate results. By combining SVM with deep learning, the accuracy of the predictions can be increased and its limitations overcome. Other factors that might impact diabetes prediction should also be considered to improve the overall accuracy of the model

## REFERENCES


[1] "Diabetes," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[2] "Saudi National Reference of Clinical Guidelines for Care of Diabetic Patients."[Online] Available: https://shc.gov.sa/Arabic/Documents/SDCP%20Guidelines.pdf.

[3] U. C. I. M. Learning, "Pima Indians Diabetes Database," Kaggle, 06-Oct-2016 [Online]. vailable:https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[4] H. El Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology", *Journal of ICT Standardization*, pp. 319–338, May 2022.

[5] U.M. Butt, et al., "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *J. Healthc. Eng.*, Vol. 2021, pp. 1–17, 2021.

[6] N. Mohan and V. Jain, "Performance analysis of support vector machine in diabetes prediction", *2020 4th International conference on electronics communication and aerospace technology (ICECA)*, pp. 1-3, 2020, November.

[7] H. Sulistyawati and A. Murtadho, "Performance Accuration Method of Machine Learning for Diabetes Prediction", *Mantik*, vol. 4, no. 1, pp. 164-171, May 2020.

[8] S. Islam Ayon and M. Milon Islam, "Diabetes Prediction: A Deep Learning Approach", *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 2, pp. 21-27, 2019

[9] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques", Proc. 1st Int. Informat. Softw. Eng. Conf. (UBMYK), pp. 1-4, Nov. 2019D.

[10] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, et al., "A model for early prediction of diabetes", *Inform. Med. Unlocked*, vol. 16, Jan. 2019.

[11] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms", *Procedia Comput. Sci.*, vol. 132, pp. 1578-1585, Jan. 2018.

[12] L. Chaves and G. Marques, "Data mining techniques for early diagnosis of diabetes: A comparative study", *Appl. Sci.*, vol. 11, no. 5, pp. 2218, Mar. 202.

[13] F. Mercaldo, V. Nardone and A. Santone, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques", *Procedia Comput. Sci.*, vol. 112, pp. 2519-2528, Sep. 2017.

[14] H. Zhou, R. Myrzashova and R. Zheng, "Diabetes prediction model based on an enhanced deep neural network", *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, pp. 148, Jul. 2020.