IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 4, June 2024

# Comparison of Machine Learning Models for Diabetes Prediction

**Dr. K. Kasturi**
Associate Professor, Department of Information Technology,
School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies, Chennai
kasturi2016research@gmail.com

**Abstract:** *The prevalence of chronic diabetic disease has significantly increased recently. Blood sugar levels rise with diabetes, which also causes additional issues like blurred vision, kidney failure, nerve damage, and stroke. Early diabetes detection helps guide the implementation of the necessary measures. Everyone's attention is being drawn to the sharp rise in the number of diabetics. Different models have been built in this study to categorize diabetic and non-diabetic individuals. The classification models for the PIMA Indian Diabetes dataset have been implemented using machine learning algorithms likeLogistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest(RF), and Support Vector Machine (SVM). Deep learning perspective algorithm such as Multi Layered Feed Forward Neural Network (MLFNN) also been implemented and comparisons were made. For better comparisons, accuracy and execution times for each algorithm are recorded. To further improve the diabetes dataset's classification accuracy, various activation functions, learning algorithms, and approaches to deal with missing information are taken into account. The results of MLFNN are then contrasted with machine learning algorithms. MLFNN has the highest achieved classification accuracy (92%) of all the classifiers and it will be more accurate if it is implemented in larger datasets. These models are built to improve the standard of the patient care. This research is helpful in predicting pre-diabetes and identifying the risk factors linked to the development of diabetes from clinical data.*

**Keywords:** Indian Diabetes data set, Machine Learning, Deep Learning, Random Forest, Support Vector Machine, Multi Layered Feed Forward Neural Network

## I. INTRODUCTION

According to data from the World Health Organization, diabetes claims the lives of almost 1.6 million people year [1].One type of sickness that arises from an extremely high blood glucose/blood sugar level in the human body is diabetes. Health professionals explain that diabetes arises from two conditions: (1) insufficient insulin production by the pancreatic gland (Type 1 diabetes) and (2) inability of body cells to utilise the insulin produced (Type 2 diabetes) [2]. Following the process of food digestion, glucose is released when we eat. A blood hormone called insulin travels from the blood to the cells, where it tells them to take up blood glucose and convert it to energy.

When the pancreas is unable to create enough insulin, glucose cannot be absorbed by the cells and stays in the circulation. As a result, blood glucose/blood sugar levels rise to an extremely unacceptably high level [3]. The human body experiences symptoms like acute hunger, severe thirst, and frequent urine as a result of elevated blood sugar. The human body typically has 70–99 mg of glucose per deciliter. Diabetes is indicated if the glucose level is greater than 126 mg/dl.

A doctor's experience and knowledge are used to predict the disease in order to diagnose patients early, yet this method is not always reliable. Therefore, the manual choices may seem concerning. Patients may not receive the right treatment as a result of decision-makers failing to recognize the hidden pattern in the data. Better automated identification accuracy is crucial for diabetes early detection.

DOI: 10.48175/IJARSCT-19072

## II. RELATED WORK

Yahyaoui, A., et.al., (2019, November)Random Forest and Support Vector Machine (SVM) (RF). Conversely, in order to forecast and identify the diabetic patients, we used a completel convolution neural network (CNN) for deep learning (DL). The public Pima Indians Diabetes database, which included 768 samples total with 8 attributes each, is used to assess the suggested approach. 268 people had diabetes, whereas 500 samples had a non-diabetic classification. DL, SVM, and RF yielded an overall accuracy of 76.81%, 65.38%, and 83.67%, respectively. The outcomes of the trial demonstrate that RF outperformed deep learning and SVM techniques in the prediction of diabetes.

Refat, M. A. R., etal., (2021, October)In order to predict diabetic disease early on, we have compared a number of ML and DL approaches in this investigation. In addition, we assessed the effectiveness of every suggested machine learning and deep learning classification algorithm using a range of performance indicators using a diabetes dataset from the UCI repository that included class.

Gupta, H.,et.al., (2022) Two prediction models using deep learning (DL) and quantum machine learning (QML) techniques have been presented based on the features found in the PIMA Indian Diabetes dataset. These generated models' predictive capacity has been assessed using the accuracy. The discriminatory performance of these models has been improved by the use of outlier rejection, normalization, and the filling in of missing variables. Furthermore, these models' performance has been contrasted with that of cutting-edge models.

Naz, H., & Ahuja, S. (2020). This study uses the PIMA data set and a variety of machine learning algorithms to propose a diabetes prediction approach.The range of 90–98% is the accuracy attained by the functional classifiers Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), and Deep Learning (DL). With an accuracy percentage of 98.07% on the PIMA data set, DL outperforms the other three in terms of diabetes onset. Therefore, this suggested system offers medical professionals an efficient prognostic tool that may aid in the early discovery of the illness.

Khanam, J. J., & Foo, S. Y. (2021) We conducted our research using the Pima Indian Diabetes (PID) dataset, which was sourced from the UCI Machine Learning Repository. The dataset includes details on 768 patients along with nine distinct attributes that correspond to each patient. We predicted diabetes using seven machine learning techniques on the dataset. We discovered that the model combining Support Vector Machine (SVM) and Logistic Regression (LR) performs well in predicting diabetes. We constructed the neural network model using distinct hidden layers across multiple epochs, and found that the NN with two hidden layers had an accuracy of 88.6%.

Saxena, R. (2021).The PIMA Indians diabetes dataset was used in this study's use of the supervised K-nearest neighbor machine learning technique. The K-nearest neighbor technique compares the similarity of data that is displayed to data that has already been stored. As a result of using the suggested algorithm, we have demonstrated that accuracy has increased by 8.48%, from 70.1% to 78.58%.

Rajendra, P., & Latifi, S. (2021).The primary algorithm in this work is logistic regression, and the Python IDE is utilized for the study. The PIMA Indians Diabetes dataset, which originated from the National Institute of Diabetes, is the primary dataset used in this project. For Dataset, the maximum accuracy of about 78% was attained.

Joshi, R. D., & Dhakal, C. K. (2021).We use a logistic regression model and a decision tree, a machine learning method, to predict type 2 diabetes in Pima Indian women in order to better understand risk variables. According to our data, age, body mass index (BMI), glucose, pregnancy, and the function of the diabetic pedigree are the five significant predictors of type 2 diabetes. We investigate a classification tree in further detail to support and confirm our findings. While the ten-node tree suggests glucose, BMI, pregnancy, diabetes pedigree function, and age as the main predictors, the six-fold classification tree reveals glucose, age, and BMI as relevant variables. Prediction accuracy of 78.26% and cross-validation error rate of 21.74% are obtained using our recommended specification.

YOU, S., & KANG, M. (2020)Utilizing Support Vector Machine (SVM), Decision Tree, and correlation analysis, we were able to identify three key variables that together account for 70% of the diabetes risk in Pima Indians. By using the information presented in this research, medical professionals may identify possible diabetic Pima Indians and prevent Indian diabetes, which has a 21.74% incidence rate.

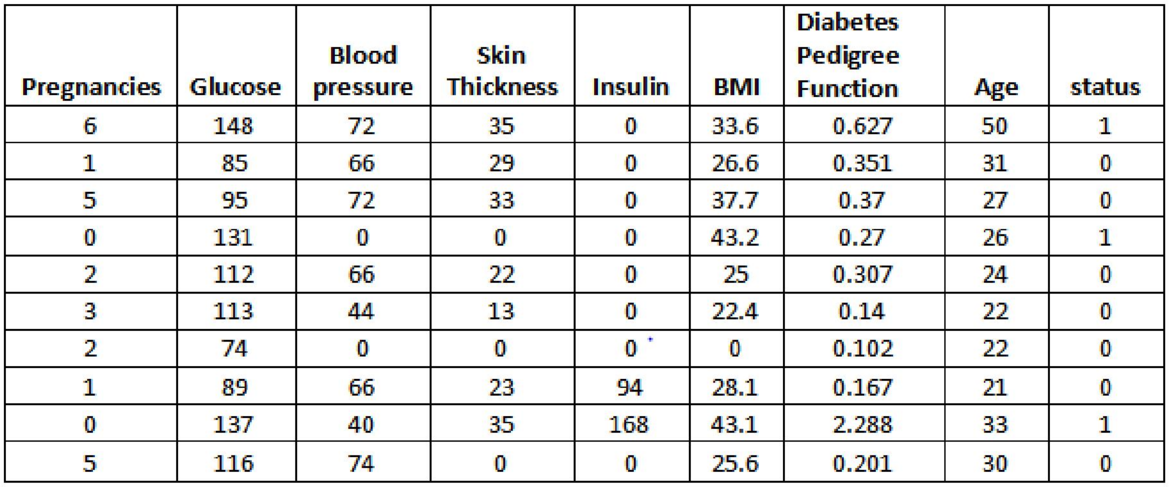
Anusha, C., et al.,(2021, August).According to this article, the Multi-Layer Perceptron (MLP) technique, which uses a multi-layer feed-forward neural network, provides the most notable expectation precision with a minimum mean square

DOI: 10.48175/IJARSCT-19072

error rate of 0.15. With an increased area under the curve of 0.88, the multi-layer perceptron (MLP) provides the lowest false negative and positive rates.

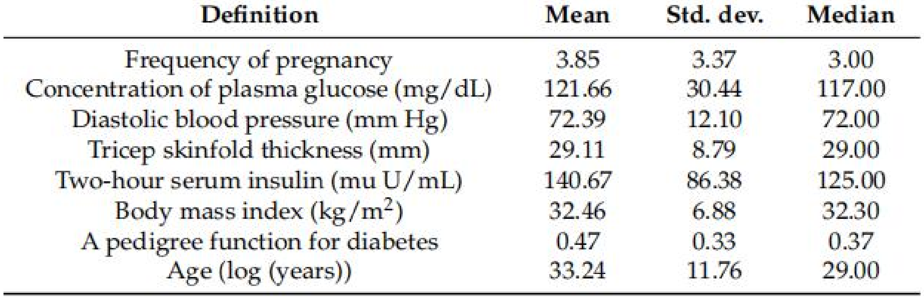
## III. DATASET

The National Institute of Diabetes and Digestive and Kidney Diseases is the original source of this dataset. Based on certain diagnostic metrics contained in the collection, the dataset aims to diagnostically predict the presence or absence of diabetes in a patient (status attribute).



| Pregnancies | Glucose | Blood pressure | Skin Thickness | Insulin | BMI | Diabetes Pedigree Function | Age | status |
|---|---|---|---|---|---|---|---|---|
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 5 | 95 | 72 | 33 | 0 | 37.7 | 0.37 | 27 | 0 |
| 0 | 131 | 0 | 0 | 0 | 43.2 | 0.27 | 26 | 1 |
| 2 | 112 | 66 | 22 | 0 | 25 | 0.307 | 24 | 0 |
| 3 | 113 | 44 | 13 | 0 | 22.4 | 0.14 | 22 | 0 |
| 2 | 74 | 0 | 0 | 0 | 0 | 0.102 | 22 | 0 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |

**Fig 3.1 Sample Data Set**



| Definition | Mean | Std. dev. | Median |
|---|---|---|---|
| Frequency of pregnancy | 3.85 | 3.37 | 3.00 |
| Concentration of plasma glucose (mg/dL) | 121.66 | 30.44 | 117.00 |
| Diastolic blood pressure (mm Hg) | 72.39 | 12.10 | 72.00 |
| Tricep skinfold thickness (mm) | 29.11 | 8.79 | 29.00 |
| Two-hour serum insulin (mu U/mL) | 140.67 | 86.38 | 125.00 |
| Body mass index (kg/m$^2$) | 32.46 | 6.88 | 32.30 |
| A pedigree function for diabetes | 0.47 | 0.33 | 0.37 |
| Age (log (years)) | 33.24 | 11.76 | 29.00 |

Fig 3.2 Mean and Std. dev of Diabetes Parameters

The selection of these examples from a wider database was subject to a number of restrictions. In specifically, every patient at this facility is a female Pima Indian who is at least 21 years old[4].

## IV. MACHINE LEARNING METHODOLOGIES

The Pima Indians onset of diabetes problem is a conventional binary classification dataset that presents an issue. There are eight numeric input variables in the problem, all with different scales, and two classes.Each algorithm is tested using the 10-fold cross validation technique, which is crucially configured with the same random seed to guarantee that the same splits to the training data are carried out and that every algorithm is evaluated exactly in the same way.

A relationship between the categorical response variable and variables is modeled by logistic regression (LR). In particular, a logistic model consists of a linear combination of independent variables with log-odds representing the probability of an event. In light of the covariate values, binary logistic regressions calculate the probability that a binary variable's characteristic is present.

A supervised machine learning approach called K-nearest neighbor (KNN) can be applied to regression and classification issues alike. Because it uses all of the data for training while classifying the data, it is often referred to as a lazy learner. It searches for features that the newly supplied data and the data that is currently available have in

DOI: 10.48175/IJARSCT-19072

common. A fresh data point is categorized according to its similarity feature. When there are precisely two classes in our data, we can use Support Vector Machines (SVM). The optimal hyperplane to divide all the data points of one class from all the data points of the other class is found by an SVM in order to classify the data. The hyperplane with the biggest margin between the two classes is the optimal one for a support vector machine.

Data that is categorized, continuous, and binary can be handled by the Random Forest (RF) Classifier. With a few restrictions, random forest is an all-around quick, easy, adaptable, and reliable model. The Random Forest algorithm is an ensemble learning method that improves a model's performance by mixing many classifiers.

The multilayer feed-forward neural network, or MLFNN, is an artificial neural network that is connected and has numerous layers. Each layer has neurons that have weights assigned to them, and the activation functions are used to compute the output. This type of network is used in deep learning. It is one of the varieties of neural networks where the network flows from input to output units; there are no loops, no feedback, and no signals that travel backward from the input layer to the hidden layer.

### 4.1 COMPARING CLASSIFICATION ALGORITHMS

The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data.

| **Algorithms** | **Accuracy** | **F1 Score** | **Precision** | **Recall** |
|---|---|---|---|---|
| **LR** | 0.73 | 0.72 | 0.72 | 0.72 |
| **KNN** | 0.75 | 0.74 | 0.74 | 0.75 |
| **SVM** | 0.76 | **0.75** | 0.76 | **0.74** |
| **RF** | 0.88 | 0.87 | 0.88 | 0.87 |
| **MLFNN** | 0.92 | 0.91 | 0.91 | 0.92 |

**Fig 4.1 Performance Metrics**

### 4.2 FEATURE IMPORTANCE

According to the Fig 4.2  the three significant factors that determine the onset of diabetes are Glucose, BMI and Age while other factors play less role in diabetes prediction.

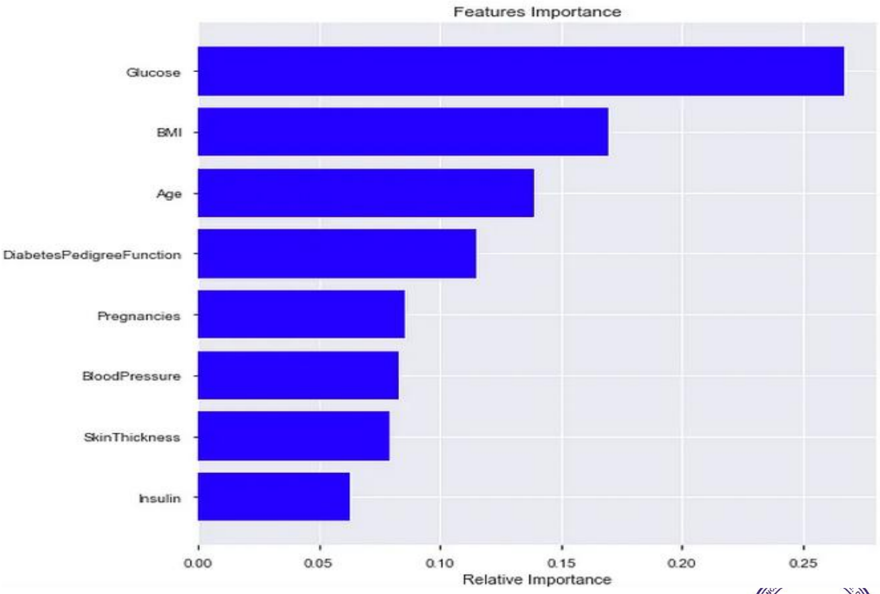


**Fig 4.2: Significant Features**

DOI: 10.48175/IJARSCT-19072

## V. RESULTS AND DISCUSSIONS

The existence of outliers contributed to the Random Forest Classifier's better performance compared to the other machine learning techniques. Outliers have less of an impact on Random Forest because it is not a distance-based approach; in contrast, distance-based algorithms like Support Vector and Logistic Regression performed worse.The best model among the machine learning algorithms is the Random Forest Classifier because of its excellent recall, accuracy, and precision scores, but compared with MLP Classifier of MLFNN the MLFNN scored high in the performance metrics.

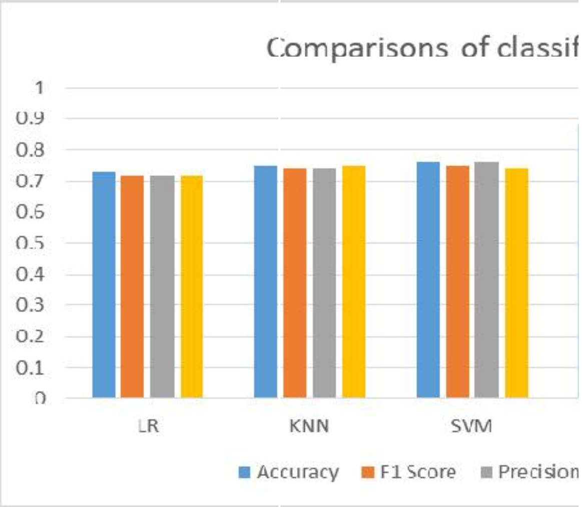


Fig 5.1: Classification comparisons

## VI. CONCLUSION

It is worthwhile to focus on many foundations in order to accurately predict and diagnose any infection through the use of machine learning. This work explores unique machine learning algorithms and their visualizations on a diabetes data set. This study examines the outcomes of machine learning computations using logistic regression, KNN, SVM, random forest, and MLP. Likewise, we can observe that the MLP has a higher accuracy of 92% compared to SVM, RF, KNN, and Logistic Regression by comparing and analyzing the results of various characterisation calculations. The results therefore suggest that diabetic individuals can be prepared and characterized using the MLP computation.

## VII. REFERENCES

**[1].** https://www.who.int/health-topics/diabetes.

**[2].** https://www.medicalnewstoday.com/articles/325018#how-is-the-pancreas-linked-with-diabetes.

**[3].** https://www.webmd.com/diabetes/diabetes-causes.

**[4].** https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

**[5].** Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International informatics and software engineering conference (UBMYK) (pp. 1-4). IEEE.

**[6].** Refat, M. A. R., Al Amin, M., Kaushal, C., Yeasmin, M. N., & Islam, M. K. (2021, October). A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) (pp. 654-659). IEEE.

**[7].** Gupta, H., Varshney, H., Sharma, T. K., Pachauri, N., & Verma, O. P. (2022). Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Complex & Intelligent Systems, 8(4), 3073-3087.

**[8].** Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19, 391-403.

**[9].** Patil, V., & Ingle, D. R. (2021, June). Comparative analysis of different ML classification algorithms with diabetes prediction through Pima Indian diabetics dataset. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-9). IEEE.

**[10].** Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. Ict Express, 7(4), 432-439.

**[11].** Saxena, R. (2021). Role of K-nearest neighbour in detection of Diabetes Mellitus. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 373-376.

**[12].** Gupta, S., Verma, H. K., & Bhardwaj, D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. In Operations Management and Systems Engineering: Select Proceedings of CPIE 2019 (pp. 365-376). Springer Singapore.

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

**[14].** Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. International journal of environmental research and public health, 18(14), 7346.

**[15].** Patil, V., & Ingle, D. R. (2021, June). Comparative analysis of different ML classification algorithms with diabetes prediction through Pima Indian diabetics dataset. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-9). IEEE.

**[16].** YOU, S., & KANG, M. (2020). A Study on Methods to Prevent Pima Indians Diabetes using SVM. Korean Journal of Artificial Intelligence, 8(2), 7-10.

**[17].** Anusha, C., Sravani, A., & Praveen, M. A. (2021, August). Diabetes Diagnosis and Classification Using Feed Forward Neural Network Algorithm. In Proceedings of the International Conference on Industrial Engineering and Operations Management (pp. 2-5).