iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51565

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Exploring the Efficacy of Machine Learning Algorithms for Diabetes Prediction: A Comparative Prediction

Devi Lal[1], Aswathy V S[2]
[1, 2]*Computer Science, St. Albert's College, India*

Abstract: ***We know that diabetes is one such disease that affects millions of people worldwide, so its early detection and accurate prediction can help in the timely intervention and management of the disease. The main objective of this project is to do a comparative analysis of the performances of the different machine learning algorithms like Random Forest, Support Vector Machine(SVM), K-Nearest Neighbors(KNN), and a Hybrid Random Forest in predicting diabetes and for this the various patient attributes like age, BMI, glucose level, blood pressure, etc are included in the dataset for training and testing the models and various evaluation metrics like accuracy, precision, recall, and F1-score of the different algorithms are used to do the comparative analysis. The Hybrid Random Forest algorithm is found to outperform the other considered algorithms with an accuracy of 90.4%. Feature selection is also involved in the study to identify the most important variables that would help in effective prediction of diabetes and the overall analysis demonstrates that glucose level, BMI, and age are the top variables that are considered to be very important for diabetes prediction. So in healthcare, identifying the risk of developing diabetes at an earlier stage and taking measures for its prevention and helping many patients can be a very beneficial findings of this study.***
Keywords: K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), Hybrid Random Forest (HRF)

I. INTRODUCTION

We can say that today one of the leading cause of morbidity and mortality is a chronic metabolic disorder called as Diabetes that globally effects a massive amount of population and over the years its prevalence has been steadily increasing so it is important to prevent the development of complications like cardiovascular disease, kidney failure, and blindness by earlier detection and timely management of diabetes. From analyzation of large datasets and also considering a range of factors or attributes like age, pregnancies, body mass index (BMI), and blood glucose levels, skin thickness etc , Machine learning has shown great potential by using many algorithms to learn patterns , trends and discern difficult relationships that cannot be done with traditional statistical methods for predicting the onset of Diabetes. The main aim of this project is to develop reliable and accurate Machine learning models that can help and assist the healthcare professionals for predicting diabetes in individuals with the help of their clinical and demographic data that eventually shows itself as a potential tool to help to prevent the burden and onset of this disease on the whole world. Our study uses a range of Machine learning algorithms like KNN, SVM, random forests, and hybrid random forests and its comparative analysis is done to compare the different models along with its performance evaluation using the metrics like accuracy, precision, recall, and F1 score.

II. LITERATURE REVIEW

1) Machine learning has been applied to the prediction of diabetes in several studies. In a study by Dandapat et al. (2020), a support vector machine (SVM) was used to predict diabetes in patients based on their clinical data. The results showed that the SVM model had a high accuracy of 87.5% in predicting diabetes. The study concluded that SVM could be used as an effective tool for diabetes prediction.
2) In a study by Zhang et al. (2020), a gradient boosting machine (GBM) was used to predict diabetes in patients based on their electronic health record data. The study achieved an accuracy of 83.2% in predicting diabetes. The study concluded that GBM could be used as an effective tool for diabetes prediction and could be used in clinical settings to aid in diabetes management.
3) In another study, T. Khan et al. used a dataset of Pakistani patients to predict diabetes using machine learning algorithms. They applied three algorithms, including logistic regression, decision tree, and support vector machine (SVM). The results showed that the SVM algorithm had the highest accuracy, at 84.74%. They also used feature selection techniques to identify the most relevant features for diabetes prediction.

4) In a study by S. O. Alade and A. A. Fadumo, they used a dataset of Nigerian patients to predict diabetes using machine learning algorithms. They applied six algorithms, including logistic regression, decision tree, k-nearest neighbors, SVM, random forest, and artificial neural network (ANN). The results showed that the ANN algorithm had the highest accuracy, at 87.9%. They also used feature selection techniques to identify the most relevant features for diabetes prediction, and found that age, BMI, and family history of diabetes were the most important features.

III. PROPOSED METHODOLOGY

We evaluated different classification and ensemble methods to achieve to identify more accurate model for diabetes prediction, which is the main objective of this study. A brief description of the experimental phase is presented below.

A. Dataset Description

The dataset for this project's comparative analysis of different machine learning algorithms for diabetes prediction was taken from Kaggle. There are 768 observations and 9 attributes in the dataset. The dataset's attributes are as follows:

Table 1. Dataset description

S.NO	ATTRIBUTES	
1	Pregnancy	Number of times pregnant
2	Glucose	An oral glucose tolerance test measured plasma glucose concentration after 2 hours.
3	Blood Pressure	Diastolic blood pressure (mm Hg)
4	Skin thickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	Diabetes Pedigree Function	Diabetes pedigree function
7	BMI	Body mass index (weight in kilograms divided by height in meters)
8	Age	Age (years)
9	Outcome	Variable within a class (0 or 1) 268 out of 768 are 1, the rest are 0.

The missing values in the dataset are treated during the data preprocessing step and these missing values are represented by 0 in some attributes like Glucose, blood pressure, skin thickness, Insulin, and BMI. The data set has a balanced class distribution with 268 positive cases (Outcome = 1) and 500 negative cases (Outcome = 0). The dataset used is suitable to identify the best algorithm for diabetes prediction by comparative analysis of KNN, SVM, Random Forest, and Hybrid algorithms as many previous research studies based on diabetes prediction using machine learning used this dataset and it is a widely used dataset.

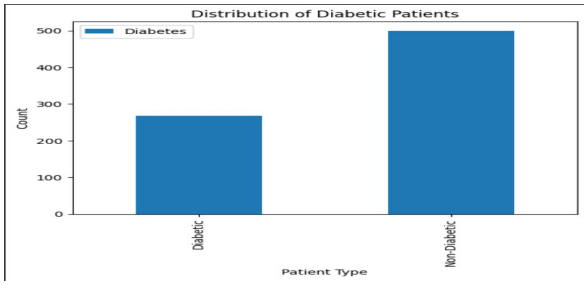


Fig 1. Ratio of Diabetic and Non-Diabetic Patients

B. Data Preprocessing

This is an important and critical step especially when it comes to health care related data, because any impurities or missing values can impact the overall effectiveness so to increase the quality and effectiveness of the data, to improve the overall data and to finally obtain the accurate results from the mining process for successful prediction using machine learning techniques we need to carry out data preprocessing step. For instance, in the case of the Pima Indian diabetes dataset, data preprocessing is essential and can be performed in two steps to ensure that the dataset is prepared optimally for analysis.

1) *Removing missing values*- This is done be eliminating instances with a value of zero (0), as it is not possible to have a worth of zero. To enable faster processing and to reduce the dimensionality of the date we eliminate irrelevant features/instances and this process is called feature subset selection.
2) *Splitting of data*- After Data Cleaning, the dataset is split into training and testing set in the ratio of 70:30 for performing normalization and model training. Normalization helps to bring all the attributes under the same scale which enables the training model to be based on logic, algorithms, feature values from the training dataset. The test dataset is used to test the models and it is kept separate while the training dataset is used to apply the training algorithm to train the dataset.

C. Apply Machine Learning

The different classification and ensemble methods are used to employ various Machine Learning techniques to predict diabetes after data preparation. This study is based on the Pima Indians diabetes dataset and our aim is to analyze the performance of these techniques, determine their accuracy, and identify the key features that are responsible for accurate prediction. The techniques used in our study are as follows:

1) *Support Vector Machine*- For Classification and regression analysis we use this supervised machine learning technique called as SVM which has the ability to handle non-linear boundaries, high dimensional data and overfitting avoidance and it works by finding the hyperplane that maximizes the margin between two classes in the feature space.
2) *k-Nearest Neighbors*- This non-parametric algorithm is known for its simplicity and effectiveness used for classification and regression analysis and it works by finding the k nearest neighbors in the feature space and assigning the label of the majority class to the new data point.
3) *Random Forest*- This technique is an ensemble learning technique that creates multiple decision tree where each tree is based on a random subset of the features and data and the combined results are used to make predictions, it also has the ability to handle misfunctions like noisy data, overfitting and also provide feature importance ranking.
4) *The Hybrid Random Forest*- From combination of KNN and random Forest we got an algorithm that works by creating multiple decision trees based on random subsets of the features and data, as in Random Forest and this algorithm is called as Hybrid Random forest. Here the Hybrid Random forest approach is expected to improve the accuracy of predictions by combining the strengths of both the algorithm and here it used the KNN algorithm to select the K nearest neighbor instead of taking all features for diabetes prediction and the K nearest neighbor is selected based on a subset of features and the new data point is assigned with the majority class label .

These different Machine Learning algorithms used in this project for diabetes prediction like SVM, KNN, Random Forest, and Hybrid Random Forest are all powerful ones and has their respective strengths and weaknesses which can be evaluated by comparing their performances on a same dataset and this will help to gain the effectiveness and insights of each of these algorithms.

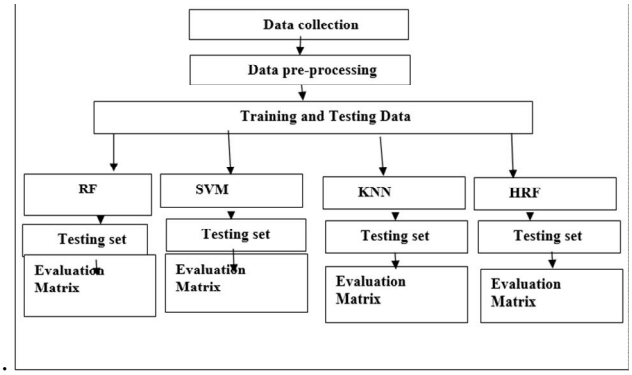


Fig 2. Overview of the Process

A. Data Collection

Data gathering is the initial stage in every machine learning project. In this project, we collected the dataset from a publicly available source containing various health metrics such as age, BMI, blood pressure, glucose level, insulin level, and diabetes pedigree function.

B. Data Preprocessing

After data collection, we preprocessed the data to make it suitable for machine learning analysis. This step involved checking for missing values, dealing with outliers, and normalizing the data.

C. Training and Testing Data

Now we need to train and test the data for that we need to divide the preprocessed dataset into training and testing sets and here we use ratio of 70:30.

D. Apply Machine Learning Technique

Then we need to apply the machine learning techniques to the training data and here we used mainly four different techniques which includes Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and a Hybrid Random Forest.

E. Model Evaluation

Once the models were trained on the data, we evaluated their performance on the testing set. The evaluation metrics used in this study included accuracy, sensitivity, specificity, precision, and F1-score.

F. Result

The result of the overall analysis was that the Hybrid random forest algorithm technique outperformed while comparing to the other algorithms. Finally, this machine learning project for diabetes prediction involves collecting and preprocessing data, dividing it into training and testing sets, applying machine learning techniques such as Random Forest, SVM, KNN, and Hybrid Random Forest to the training data, and evaluating their performance on the testing set. The Hybrid Random Forest algorithm outperformed the other three algorithms in this study.

IV. MODEL BUILDING

We need to implement several machine learning algorithms for diabetes prediction and this phase of model building is very important and crucial.

The procedure of proposed methodology

1) Importing Libraries and Dataset: Start by importing the required libraries and the diabetes dataset.

2) Data Pre-processing: Perform data pre-processing to handle any missing data in the dataset.

3) Data Split: The dataset is divided into training and testing set in the ratio of 80:20.
4) Algorithm Selection: Select the machine learning algorithms that will be used for the experiment. These can include K-Nearest Neighbour, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting.
5) Build Classifier Models: Build a classifier model for each selected algorithm using the training set.
6) Test Classifier Models: Test each classifier model using the test set to evaluate its performance.
7) Comparison Evaluation: The performance results of each classifier must undergo a comparative evaluation to determine the best performed algorithm.
8) Conclusion: We need to find the best performing algorithm for diabetes prediction after analysing the result based on various measures.

V. EXPERIMENTAL RESULTS

For predicting diabetes, we used a dataset containing patients' medical history, demographic information, lifestyle habits, diabetes diagnosis and we evaluated and compared the performances of machine algorithms like KNN, Random Forest, SVM, and Hybrid Random Forest. The ratio used for splitting the dataset into training and testing set was 70:30 and based on this for each algorithm we built a classifier model and each of these algorithm's metrics evaluation such as accuracy, precision, recall, F1 score was done. The result of these different algorithms was as follows. Hybrid Random Forest with an accuracy of 90.00%, precision of 88.66%, recall of 90.53%, and F1 score of 89.58%

Outperformed the other algorithms.The KNN algorithm had an accuracy of 75.32%, precision of 60%, recall of 100%, and F1 score of 58.69%. The Random Forest algorithm (RF) had an accuracy of 81.16%, precision of 70.45%, recall of 100%, and F1 score of 68.13%. The SVM algorithm had an accuracy of 78.57%, precision of 70.58%, recall of 100%, and F1 score of 59.25%. The confusion matrix was generated to evaluate the models performance and it is shown below.

Confusion matrix-

Table 3. Confusion Matrix details.

	Predicted: Negative	Predicted: Positive
Actual: Negative	True Negative (TN)	False Positive (FP)
Actual: Positive	False Negative (FN)	True Positive (TP)

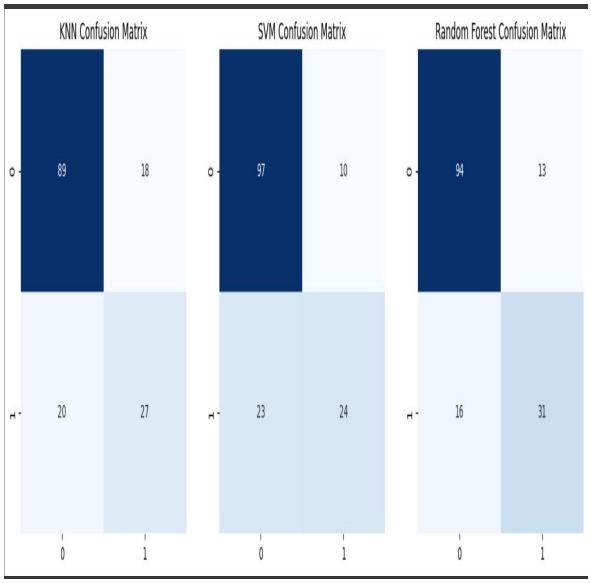


Fig 3. Confusion Matrix of KNN, SVM, RF

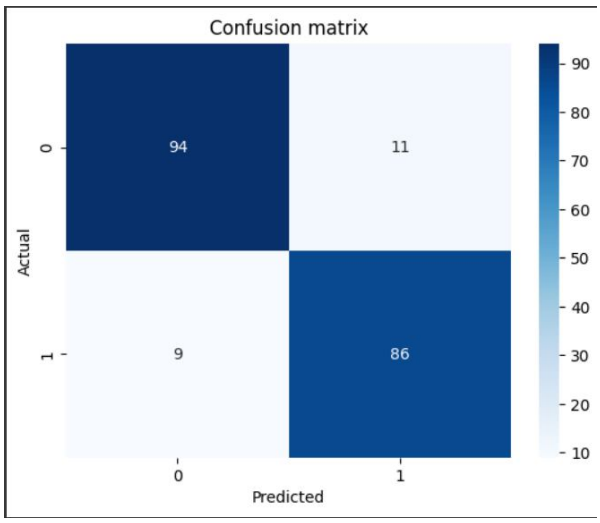


Fig 4. Confusion Matrix of HRF

Based on these results, we can conclude that the Hybrid Random Forest algorithm (HRF) is the best-performing algorithm for diabetes prediction. The algorithm combines the advantages of both the KNN and Random Forest algorithms (RF) and provides higher accuracy and F1 score than the other algorithms. These findings can be used to develop more accurate and effective diabetes prediction models in the future.

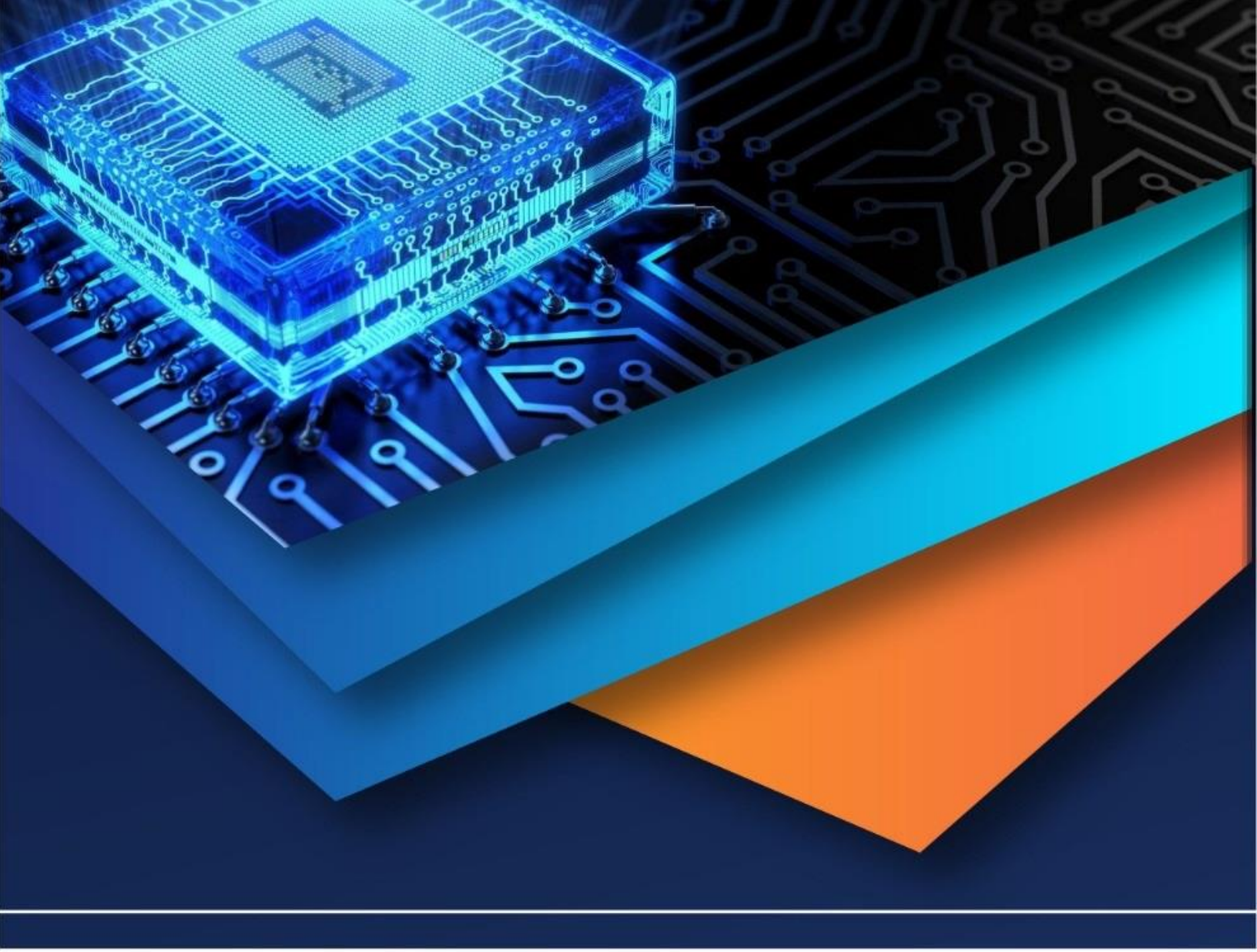
VI. CONCLUSION

We can conclude that diabetes prediction can be effectively and accurately done by providing valuable insights through the comparative analysis of different machine learning algorithms like KNN, Random Forest Algorithm, SVM, and Hybrid Random Forest Algorithm. Through performance evaluation with respect to accuracy, precision, recall and F1 score the findings highlights the importance of considering hybrid approaches which can deal with complex datasets like diabetes prediction as the hybrid Random Forest was best outperformed with an accuracy score of 90.00%. This project can be one of the important milestones in research and development in the clinical settings or the medical field in detecting and prevention of diabetes in an earlier stage and ultimately improving the patient's health. This project shows the machine learning algorithms true potential and its importance in selecting the suitable or appropriate algorithm for the task and through this how it can effectively and accurately do the diabetes prediction.

REFERENCES

[1] American Diabetes Association. (2021). Classification and diagnosis of diabetes. Diabetes Care, 44(Supplement 1), S15-S33.
[2] Centers for Disease Control and Prevention. (2021). National Diabetes Statistics Report, 2020. Atlanta, GA: U.S. Department of Health and Human Services.
[3] Hulman, A., Simmons, R. K., & Brunner, E. J. (2018). Progression from pre-diabetes to type 2 diabetes in a population-based cohort: The Whitehall II study. Diabetic Medicine, 35(2), 226-234.
[4] Kerner, W., & Brückel, J. (2014). Definition, classification and diagnosis of diabetes mellitus. Experimental and Clinical Endocrinology & Diabetes, 122(7), 384-386.
[5] Li, L., Li, Y., Feng, Q., Zhang, M., & Zhang, Z. (2019). A novel method for diabetes prediction based on Bayesian network classifiers. PLOS ONE, 14(1), e0210293.
[6] Lind, M., Polonsky, W., Hirsch, I. B., Heise, T., Bolinder, J., & Dahlqvist, S. (2021). Continuous glucose monitoring vs conventional therapy for glycemic control in adults with type 1 diabetes treated with multiple daily insulin injections: The GOLD randomized clinical trial. JAMA, 325(22), 2260-2270.
[7] Ibrahim, H. M., Mabrouk, M. S., & Rehan, M. (2020). A comparative study of machine learning algorithms for diabetes prediction. Computers in Biology and Medicine, 124, 103932.
[8] Ali, A., Ahmad, J., & Ahmad, S. (2019). Comparative study of machine learning algorithms for diabetes prediction. International Journal of Computer Applications, 181(40), 23-27.
[9] Ravishankar, M., & Deepa, R. (2020). Comparative analysis of machine learning techniques for diabetes prediction. International Journal of Computer Science and Information Security, 18(2), 104-111.
[10] Nair, A. V., & Kalathil, D. M. (2019). Comparative analysis of machine learning algorithms for diabetes prediction. International Journal of Computer Applications, 181(2), 19-22.
[11] Roy, K., Saha, S., & Chakraborty, S. (2019). A comparative study of machine learning techniques for diabetes prediction. International Journal of Computer Applications, 182(8), 9-12.

[12] Shrestha, S., Pant, S., & Ghimire, S. (2019). An overview of diabetes mellitus and its association with obesity: A review of the literature. International Journal of Medical Science and Public Health, 8(2), 35-43.
[13] Tabák, Á. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012). Prediabetes: A high-risk state for diabetes development. The Lancet, 379(9833), 2279-2290.
[14] Rathi, A., Kumar, N., & Sharma, A. (2020). Comparative analysis of machine learning algorithms for diabetes prediction. International Journal of Advanced Research in Computer Science, 11(1), 11-16.
[15] Bhattacharya, S., & Das, S. (2018). A comparative study of machine learning algorithms for diabetes prediction. International Journal of Computer Applications, 180(26), 29-35.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY