2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-6654-9400-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAECT57570.2023.10118315

Comparison of Performance of Machine Learning Algorithms for Diabetes Detection

Pranav Dalve
Department of Electronics and Telecommunication Engineering,
Vishwakarma Institute of Technology
Pune , India.
ORCID ID : 0000-0001-8615-1026
Email: pranav.dalve20@vit.edu

Dyuti Bobby
Department of Electronics and Telecommunication Engineering,
Vishwakarma Institute of Technology
Pune , India.
ORCID ID: 0000-0001-5647-7383
Email: dyuti.bobby20@vit.edu

Abha Marathe
Department of Electronics and Telecommunication Engineering,
Vishwakarma Institute of Technology
Pune , India.
ORCID ID : 0000-0002-8457-4305
Email: abha.marathe@vit.edu

Atharva Dusane
Department of Electronics and Telecommunication Engineering,
Vishwakarma Institute of Technology
Pune , India.
ORCID ID : 0000-0003-1693-3457
Email: atharva.dusane20@vit.edu

Shreya Daga
Department of Electronics and Telecommunication Engineering,
Vishwakarma Institute of Technology
Pune , India.
ORCID ID :0000-0002-0153-0033
Email: shreya.daga20@vit.edu

Abstract*— In *today's society, Diabetes Mellitus affects a large portion of the population. The purpose of this research is to examine a few Machine Learning (ML) algorithms to assist in prediction of Type 2 diabetes, a disorder that alters the body's blood sugar processing. For the diabetes prediction proposed, the prominent PIMA Indian Diabetes Dataset was utilized. Support Vector Machine (SVM), K-Nearest Neighbours (KNN), XGBoost (Extreme gradient boosting), and Random Forest Regression are the techniques that were investigated. SVM yields an accuracy of 84.5%, sensitivity of 88.00% and specificity of 81.00%. KNN yields an accuracy of 80.5% sensitivity of 83.00% and specificity of 78.00%. XGBoost yields an accuracy of 83.41%, sensitivity of 80.18% and specificity of 86.61%. Random Forest Regression yields an accuracy of 91.26%, sensitivity of 82.76% and specificity of 98.96%. Diabetes may be detected early and treated promptly, slowing the course of the disease. Using an ML model to reliably foresee Type 2 diabetes might prove beneficial. As a result, machine learning and artificial intelligence have found a home in the healthcare industry. The Random Forest method should be used for a far more accurate and specific identification, according to a comparison of the models.

Keywords- Diabetes detection, KNN, Machine learning, Random Forest, Support Vector Machine, XGBoost

I. INTRODUCTION

Diabetes Mellitus, commonly called diabetes, is a group of diseases plaguing more than 422 million people worldwide, as asserted by the World Health Organization (WHO) [1], with more than 95% of the affected suffering from Type 2 diabetes. It occurs as a result of the inability of the human body to process blood sugar. According to a survey, more than 69.2 million people have type two diabetes. India has the world's second-highest number of individuals living with diabetes mellitus. The loss of pancreatic beta cells, which leads insulin insufficiency, is the primary cause of this illness. Insulin inadequacy is a result of the ineffective delivery of insulin to the cells [2]. If this disease is not treated in a timely manner, a person can develop serious health problems such as diabetic ketoacidosis, hyperosmolar hyperglycemia, and even death. This can lead to lifelong complications such as cardiovascular disorder, stroke, kidney collapse, foot, and eye ulcer complications, etc. [3]. As a result, early identification of diabetes is critical for intervening at the appropriate moment to prevent the disease's progression and consequences [4].

The organization of the remaining paper begins with the literature review, followed by the methods and materials, results and discussions and finally the conclusion.

II. LITERATURE REVIEW

In the domain of diabetes detection, there are different models which represent different accuracies. One of the papers proposes DLPD (Deep Learning for Predicting Diabetes) model to predict the occurrence of diabetes along with a classification of the type. They have also used gradient descent in their model [2]. A similar conclusion was established by using multiple algorithms in which Random Forest performed the best and performed well even in the external validation processes [3]. On the other hand, it was found that the Deep Learning model has the greatest accuracy as compared to all the other papers in the literature survey. Hyperparameter tuning was carried out before training the model, which gave a very high accuracy [4,12]. Marked by high blood glucose levels over a prolonged duration of time, its symptoms include increased thirst, frequent fatigue, unexplained weight loss and delayed wound healing [5]. If timely treatment is avoided, diabetes can also result in serious health conditions like cardiac diseases, diabetic ketoacidosis, etc. Diabetes detection is a field of study that has been widely explored over the years. One of the studied papers has used SVM and Random Forest and analyzed the Principal Component

Analysis dimensionality reduction method to identify the chances of diabetes and diabetes inflicted diseases. The authors have performed dimensionality reduction and feature selection in data pre-processing. The authors also concluded that dimensionality reduction is not of any high relevance [6]. Diabetes can also be predicted using the Backpropagation algorithm, J48, naive Bayes, and SVM on the dataset. The backpropagation model gave better accuracy than the other one [7]. Decision trees, Naïve Bayes, and Random Forest algorithms have been deployed while the best results were achieved in the Random Forest model which performed better than the rest models [8]. A proposed multilayer perceptron model with radial basis function (RBF) and stochastic gradient descent can be used for removing misclassified instances, achieving a significant accuracy [9]. While comparing results on a different dataset, similar results were found, and Random Forest gave the best accuracy amongst all models [10]. While deploying Linear SVM, Radial SVM, KNN, and Neural Networks. The linear SVM gave the finest accuracy as evaluated among all the models including the neural networks [11]. Improved K-means and Logistic Regression were the algorithms applied to achieve an increase in accuracy by 3.04% [13]. Globally, researchers have tried a variety of ML algorithms to aid in the convenient sensing of insulin-dependent diabetes [14-16].

Analyzing different research techniques deployed, this paper proposes a few ML algorithms, to increase the accuracy of existing systems and investigate superior techniques for the purpose of diabetes detection. The proposed algorithms are SVM, KNN, XGBoost and RF.

III. METHODS AND MATERIALS

A. Dataset

The PIMA India Dataset [17] is used in this project, originally from the National Institute of Diabetes and Digestive and Kidney Disease. All patients considered are females, above 21 years of age residing in Phoenix, Arizona, USA. The dataset includes several medical predictor variables like Diastolic blood pressure, Plasma glucose, pregnancies, Diabetes pedigree function, Insulin etc., along with the single target variable "Outcome".

In a nutshell, the dataset consists of nine columns with eight predictors and one outcome parameter that clearly identifies 768 observations of 268 positive for diabetes and 500 negatives for diabetes. The outcome variable has two classes, class 1 indicates people with diabetes whereas class 0 indicates others.

TABLE 1. Description of dataset features

Sr no.	**Feature name**	**Description**
1.	Pregnancies	Number of pregnancies undergone by patient
2.	Glucose	Plasma glucose concentration after two hours in an oral glucose tolerance test
3.	BloodPressure	Diastolic blood pressure in mmHg
4.	SkinThickness	Triceps' skin fold thickness in mm
5.	Insulin	2-Hour serum Insulin in mu U/ml
6.	BMI	Patient Body Mass Index in kg/m^2
7.	Diabetes Pedigree Function	Likelihood of diabetes attributed to family history
8.	Age	Patient age in years
9.	Outcome	Class variable: 0- unaffected, 1- affected

B. Data Analysis

The following results can be inferred after analysing the dataset,

1. The maximum number of women included in the dataset have been pregnant 0-5 times.
2. The maximum number of dataset entries indicate blood glucose levels to be around 100 mg/dl or higher, but fewer more than 150 mg/dl.
3. The body mass index is considered a global measure of health. In the PIMA Indians Dataset, the maximum percentage of people have body mass index (BMI) between 30 and 40.
4. Diabetes pedigree function is a score of likelihood of diabetes in people, based on the family history. Majority of the values here lie between pedigree function 0 and 0.5.
5. The PIMA Indian dataset includes women above 21 years of age. A large percentage of the women included are in their early 20s.
6. The likelihood of having Diabetes is highest when BMI (Body Mass Index) is highest, with a peak in age of around 40 years.
7. The occurrence of diabetes as well as glucose levels increase with an increase in age.

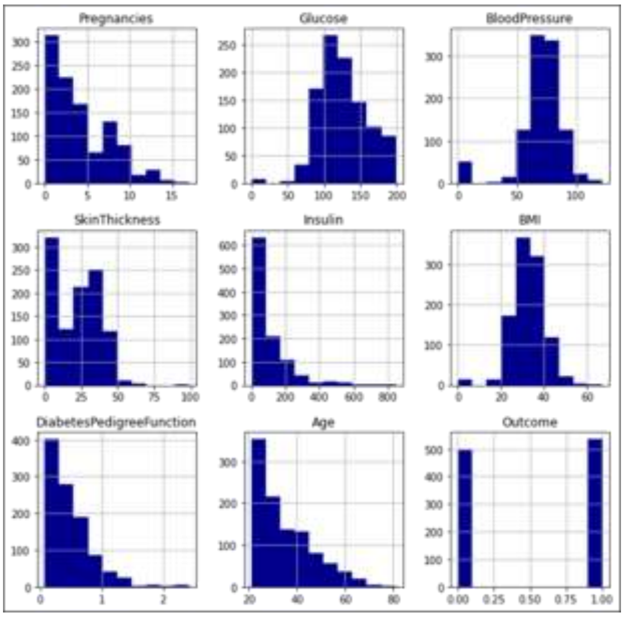


Figure 1: Relation between variables and outcomes

C. Data Preprocessing

Initially the dataset was scoured for missing data. Preprocessing for the same was not required as no instances of missing data were found in the data used. Then classifiers such as Random Forest, XGBoost, SVM and KNN were used for classification of the dataset into Class 0 – unaffected and Class 1 – affected.

There were 768 values in the initial dataset, including 268 persons with diabetes (Class-1) and 500 people without diabetes (Class-0). Naturally, the model is biased towards the class with the larger number of training values, thereby decreasing its actual prediction accuracy.

To solve the issue of imbalance in the data, up sampling is performed on training data for each model. In the process of up sampling, data points corresponding to the minority class are synthetically generated and added to the working data, by sampling with replacement. This ensures that there are the same number of counts for each label.

The dataset was split in the ratio of 80:20 where 80% of the data was used to prepare the training part for the model whereas the latter half was used for the testing purpose.

On the processed data, the further listed algorithms were applied and tested.

D. Algorithms used

a) SVM:

SVM is a supervised machine learning (ML) model. It has the objective to find a hyperplane in an n-dimensional place, such that maximum distance exists between data points of distinct classes. The purpose of maximizing this distance(margin) is to reinforce future classifications.
The kernel is a collection of mathematical functions used in SVM algorithms. The kernel's job is to take data and convert it into the appropriate format. There are various types of functions that can be used. Examples include linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid functions. The first kernel tried for the proposed work was Linear Kernel which has been mathematically depicted in Eq. (1).

$$k(x,y) = x^T.y \quad (1)$$

A significant hike in accuracy (10.69%) was observed for Gaussian Radial Basis function (RBF) which has been mathematically depicted in Eq. (2).

$$\emptyset(x) = \exp\left(\frac{-x^2}{2\sigma^2}\right), \sigma > 0 \quad (2)$$

Preprocessing was used to center, scale, and convert the data, and ROC was used as the metric.
The tuning parameter “sigma” for Radial-SVM was maintained at 0.1347335. Using 5-fold cross validation and evaluating the Receiver Operating Characteristic (ROC), an optimal model was selected with C (a regularization parameter) = 4, as seen in Figure 2.

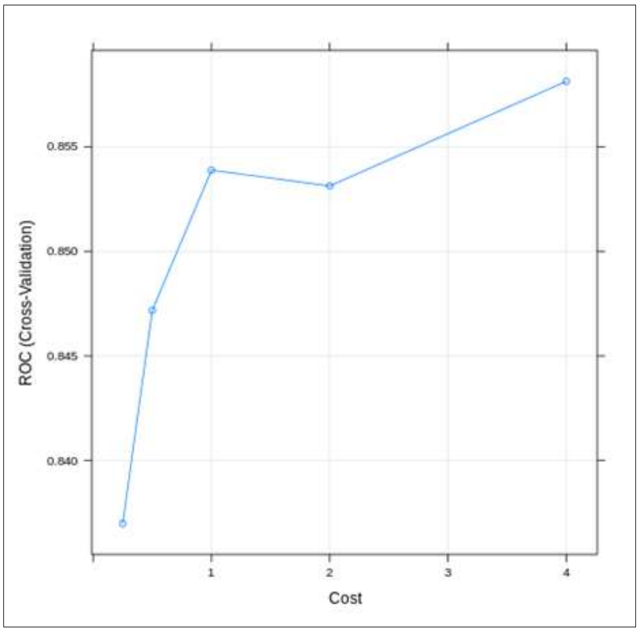


Figure 2: SVM ROC for different cost values

It has a final accuracy of 84.5%, sensitivity of 88% and specificity of 81%. The f1- score calculated based on the SVM model validation results is 0.8514.

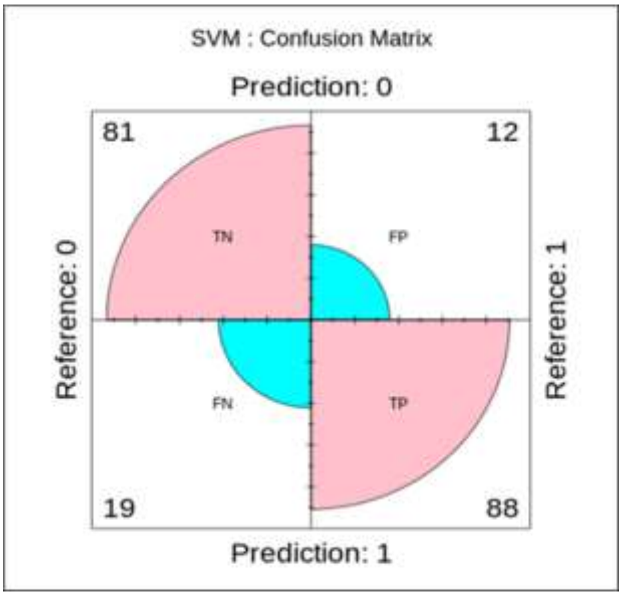


Figure 3. Confusion matrix for SVM model

b) K- Nearest Neighbors (KNN):

KNN is another supervised ML model, which focuses on most of the classes of the K – nearest neighbors of the sample data for its classification. Preprocessing was used to center, scale, and convert the data, and ROC was used as the metric.

Using 10-fold cross validation and evaluating the ROC, an optimal KNN model was selected with K (number of nearest neighbors to be considered) = 5.

The ROC increased from smaller values of k, peaked at ‘5’ and started decreasing as we further moved to higher k values as clearly depicted in Figure 4. Hence the optimum k value, chosen was ‘5’.

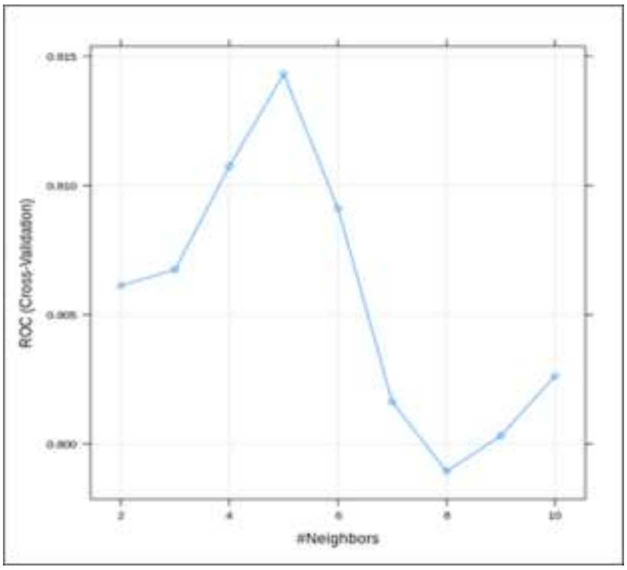


Figure 4. KNN model characteristics

The achieved accuracy when k = 6 is 80.5%, sensitivity is 83% and specificity is 78%. The f1-score calculated is 0.8097.

The confusion matrix for the above model is depicted in Figure 5.

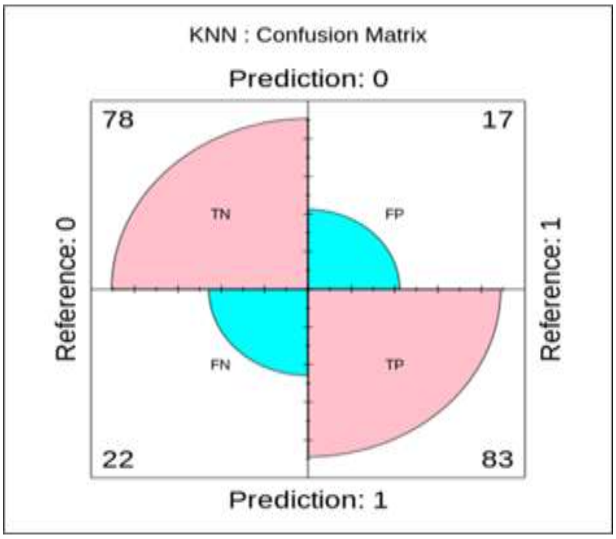


Figure 5. Confusion matrix for KNN model

c) XGBoost:

XGBoost, also known as Extreme gradient boosting, is an implementation of decision trees which have been gradient boosted. It functions as a technique to give optimized results in the shortest possible time.

For deploying the XGBoost algorithm, the data has been transformed into a matrix. The model used for maximum accuracy has an ETA (learning rate) of 0.07, with the number of decision trees as 50. The confusion matrix for the XGBoost model is depicted in Figure 6.

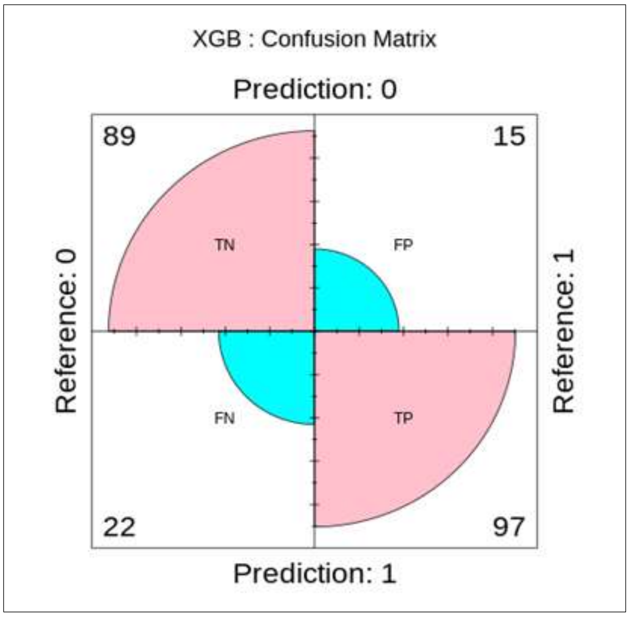


Figure 6. Confusion matrix for XGBoost model

The n-thread parameter was set to -1 so that the model could utilize all the CPU cores. For training the model, the objective function was binary logistic.

The model has a final accuracy of 83.41%, sensitivity 80.18% and specificity 86.61%. The calculated f1-score is 0.8312.

d) Random Forest (RF):

RF is an ensemble of decision tree models focusing on the concept of bagging and feature randomness. Bagging of specific features is performed to receive an accurate and more stable prediction. A major advantage of RF over decision trees is in the prevention of overfitting of the model.

Figure 7 depicts the decrease in error with an increase in the number of trees ranging from 0 to 550. Trying the model for a lesser number of trees resulted in a less accurate model while 550 trees were enough to give required accuracy.

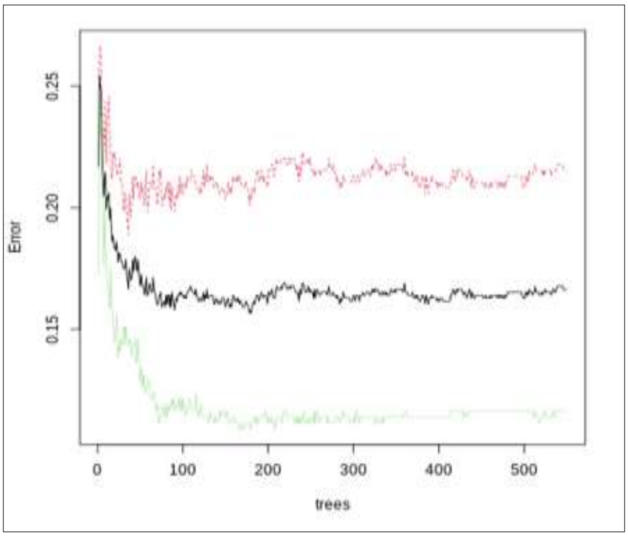


Figure 7. RF model characteristics

The Out of Bag error (OOB) was the least (16.16%) for the value of m-try being '2', while it being the highest (16.52%) for m-try greater than equal to '8', as depicted in Figure 8.
The m-try and n-tree values chosen were 2 and 550, respectively.

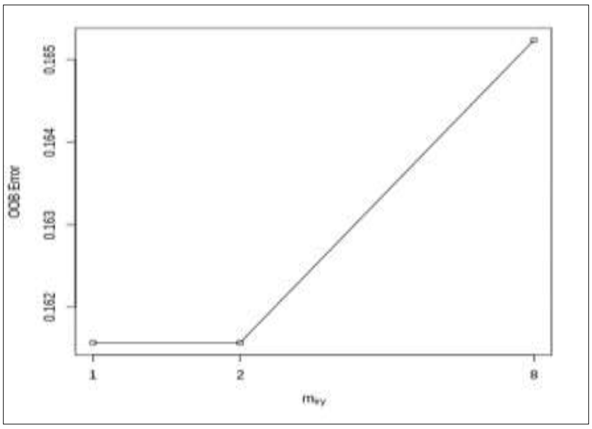


Figure 8. Tuning of Random Forest

The confusion matrix of the above stated model has been described in Figure 9.

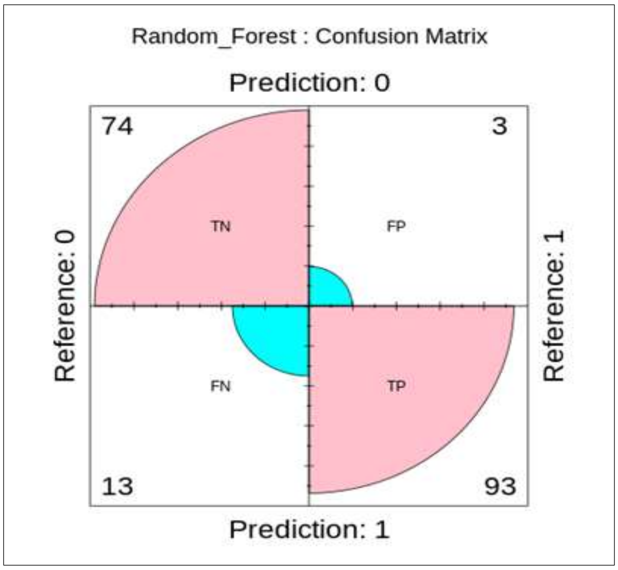


Figure 9. Confusion matrix for Random Forest model

The Random Forest model, with an accuracy of 90.71%, sensitivity of 85.06% and specificity of 95.83%. The calculated f1-score of the RF model is 0.897.

IV. RESULTS AND DISCUSSIONS

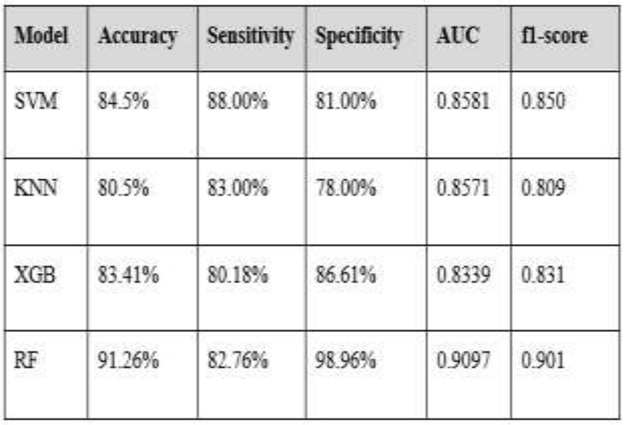
The true positive rate or the 'Recall' was achieved the maximum in SVM followed by RF and XGB models respectively. In contrast, the true negative ratio soared high in case of RF followed by XGB and SVM models respectively.

The f1 score also followed a similar trend as compared to the specificity, with the maximum f1 score for RF model which indicates the best accuracy on test dataset.

The Area Under Curve (AUC) peaked in the RF model (0.9097) followed by SVM model (0.8581), KNN (0.8571) and XGB model (0.8339). This clearly states that the Random Forest model performed better at distinguishing the positive and negative classes.

On comparing the various models used, Random Forest emerged as the best algorithm for diabetes detection.
Table 2 depicts the analogy of all four models along with the comparison parameters as Accuracy, Sensitivity, Specificity, AUC and f1 score.

TABLE 1. Results of models applied



Model	Accuracy	Sensitivity	Specificity	AUC	f1-score
SVM	84.5%	88.00%	81.00%	0.8581	0.850
KNN	80.5%	83.00%	78.00%	0.8571	0.809
XGB	83.41%	80.18%	86.61%	0.8339	0.831
RF	91.26%	82.76%	98.96%	0.9097	0.901

The above table can be visualized using a bar chart as shown in Figure 10.

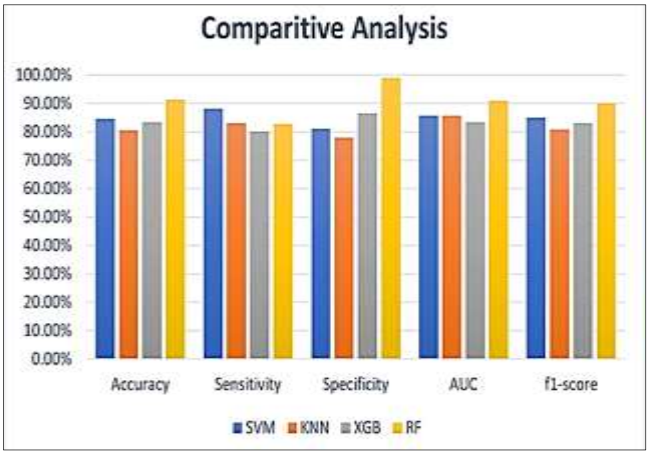


Figure 10. Comparative analysis of algorithms

The ROC/ AUC curves are plotted jointly to get a better evaluation of all models used as seen in Figure 11.

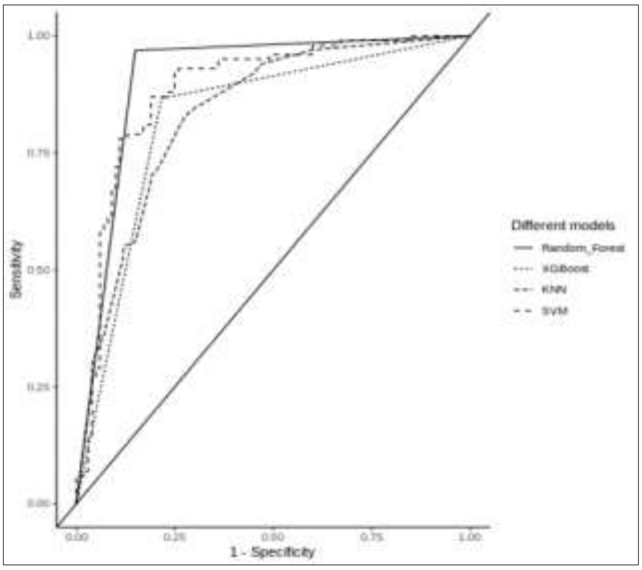


Figure 11. AUC comparison

V. CONCLUSION

To conclude, diabetes mellitus is a condition that affects both seniors and young persons in today's society. Cure for the same is available, but timely detection and prediction of diabetes can go a long way in preventing serious complications. Therefore, machine learning and artificial intelligence have found their place in healthcare. The proposed system too aims to be a positive step in this direction, testing out different algorithms to tackle this global issue. Comparison of the models yields that Random Forest algorithm should be the one preferred for a much more accurate and specific detection. As a part of the future scope, the model can be deployed on a website. Increased accuracy models can also be used to predict the risk of specific diabetes inflicted diseases. A wider dataset based on real-time clinical results can be incorporated. External validation can be done using various data sources.

REFERENCES

[1] World Health Organization (WHO), November 2021. [Cited on: 29/01/2022]. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Zhou Huaping, Raushan Myrzashova, and Rui Zheng. Diabetes prediction model based on an enhanced deep neural network. EURASIP Journal on Wireless Communications and Networking 2020, no. 1 (2020): 1-13.

[3] Tigga Neha Prerna, and Shruti Garg. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science 167 (2020): 706-716.

[4] Naz Huma, and Sachin Ahuja. Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders 19, no. 1 (2020): 391-403.

[5] Ramachandran A. Know the signs and symptoms of diabetes. The Indian journal of medical research 140, no. 5 (2014): 579.

[6] Sivaranjani S., S. Ananya, J. Aravinth, and R. Karthika. Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 141-146. IEEE, 2021.

[7] Woldemichael Fikirte Girma, and Sumitra Menaria. Prediction of diabetes using data mining techniques. 2nd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 414-418. IEEE, 2018.

[8] Yuvaraj N., and K. R. SriPreethaa. Diabetes prediction in healthcare systems using machine learning

algorithms on Hadoop cluster. Cluster Computing 22, no. 1 (2019): 1-9.

[9] Ranjeeth S., and Venkata Ajay Krishna Kandimalla. Predicting Diabetes Using Outlier Detection and Multilayer Perceptron with Optimal Stochastic Gradient Descent. 2020 IEEE India Council International Subsections Conference (INDISCON), pp. 51-56. IEEE, 2020.

[10] Katarya Rahul, and Sajal Jain. Comparison of Different Machine Learning Models for diabetes detection. 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), pp. 1-5. IEEE, 2020.

[11] Kaur Harleen, and Vinita Kumari. Predictive modeling and analytics for diabetes using a machine learning approach. Applied computing and informatics (2020).

[12] Gupta Himanshu, Hirdesh Varshney, Tarun Kumar Sharma, Nikhil Pachauri, and Om Prakash Verma. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Complex & Intelligent Systems (2021): 1-15.

[13] Wu Han, Shengqi Yang, Zhangqin Huang, Jian He, and Xiaoyi Wang. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked 10 (2018): 100-107.

[14] Islam Md Merajul, Md Jahanur Rahman, Dulal Chandra Roy, and Md Maniruzzaman. Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, no. 3 (2020): 217-219.

[15] Kavakiotis Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas et al. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal 15 (2017): 104-116.

[16] Nirmala Devi M., S. Appavu alias Balamurugan, and U. V. Swathi. An amalgam KNN to predict diabetes mellitus. 2013 IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN), pp. 691-695. IEEE, 2013.

[17] PIMA Indians Diabetes Dataset – UCI Machine learning. Available from: https://www.kaggle.com/uciml/pima-indians-diabetes-database [Cited on 29/01/2022]