A COMPARATIVE STUDY ON MACHINE LEARNING ALGORITHMS FOR DIABETICS MELLITUS

Dr.S.Manju[1]
Associate Professor, Department of Master of computer Applications, PSG College of Arts & Science Coimbatore, India
A.Jaishree and L.Harini[2]
PG Students, Department of Master of computer Applications,, PSG College of Arts & Science Coimbatore, India
N.Thamaraikannan[3]
Ph.D Research Scholar, PG and Research Department of Computer Science, PSG College of Arts and Science, Coimbatore India

Abstract - **Diabetes is a disease that affects how your body processes blood sugar and is often referred to as diabetes mellitus. In future the world's diabetic patient will reach 642 billion which means that one of the 10 adults in the future is suffering from diabetics The Objective of this paper is to made a comparative study on diabetic's mellitus Prediction. In this study we compared Machine Learning Algorithms Likely Decision Tree, Random Forest, Neural Network, Support Vector Machine [SVM], Naive Bayes, KNearestNeighbor, Logistic Regression, Gradient boost, Multilayer Perceptron, Adaboostalgorithm, Principal Component Analysis [PCA]. Statistics from 2019 showed that persons who were 18 years of age or older had diabetes, and figures from 2022 show that the disease alone was responsible for 1.5 million fatalities. The study results when compared with previous research shows the better algorithm which gives the accurate results on clinical dataset. This discovery has great impact on clinical practice. This aims at diagnosing diabetics' disease of health workers at its early stage. This result helps the researchers to build a better proposed model.**

Keywords: Diabetics, Machine Learning, Decision Tree, Naive Bayes, Support Vector Machine, Random Forest.

I. I.INTRODUCTION

Diabetics Mellitus means high level of blood glucose which causes severe damage of the heart, blood vessels, eyes, kidneys and nerves. In India the prevalence of diabetics has considerably increased as well current research reveals that out of 1, 00,000 people by 2040.124874.7[9]. To achieve this, this work traverse's traditional diabetics by taking variate attributes related to diabetic's disease. For this purpose, prediction model with a high accuracy of 95.4% which classify all the majority classes properly while mislabelled all the minority classes. We have comparatively discussed the model accuracy. The existing research discusses earlier significant work on early prediction based on Machine Learning Techniques. In this study we have compared different datasets, parameters and various Machine Learning Algorithms to find the better Algorithm to give the accurate results on the given dataset. This result helps the healthcare centers to early prevention of diabetics. Hence in this survey, attempts were taken to review the current literature on Machine Learning and data mining approaches in diabetic's research. The main objective of this study is to analyze the possibility of diabetics at early stage in order to obtain better results which compared to previous research study. Optimistic results can be used for further study on this topic.

II. LITERATURE REVIEW

Early diabetics can be diagnosed while the disease in its early stages. The Algorithm is applied though it can classify the risk of diabetic's mellitus. The majority of the concerned literature makes you of the Pima Indian diabetes dataset as it's informative index. Diagnosis of diabetes is a growing area of study. Dataset was first collected and validated then normalized to achieve numerical stability before pre-processing operations could be performed. This study contains a summary of the works suggested by different researchers in the field of diabetes. In the recent times it has always been the developing, dependable and supportive technology in the medical sector. Diabetes mellitus is amongst the most significant severe problems in the medical profession. In the purpose of this comparative analysis is to assess the classifiers that can predict the probability of disease in patients with the greatest precision and accuracy. Based on precision and accuracy we are coming to the conclusion that which algorithms suits best for the research. The precision and accuracy rate varies based on the parameters used in the research. In

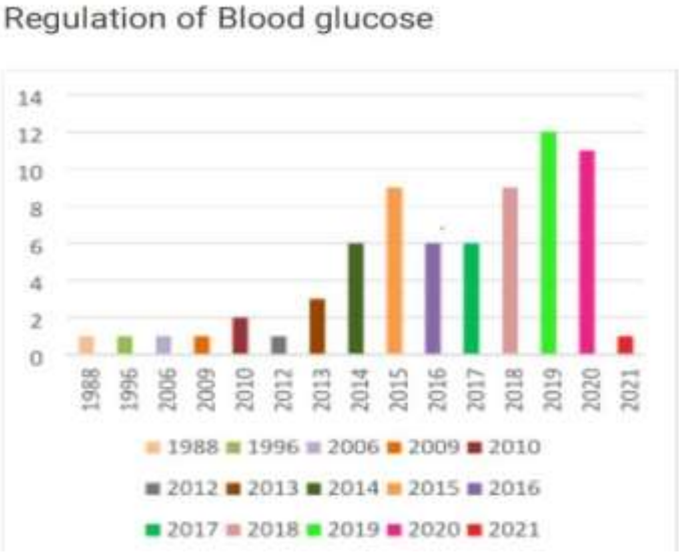
order to provide a comprehensive classification and comparison of existing techniques using key parameters and to highlight the corresponding challenges in the field of diabetes prediction. The above algorithms applied which yields the best outcomes of all algorithms and tested the results of the research [5]. We have analyzed the research papers from [2016-2022]. While comparing the previous papers we come to know that support vector machine gives the best result but the Random Forest Algorithm gives the most accurate results in many of the cases. Random Forest Algorithm utilizes Machine Learning Technique as the classifier for analysis of diabetics. Random Forest Algorithm is a process that operates among multiple decision trees to get the optimum result by choosing the majority among them as the best value [11]. Thus, it's developing a system which can predict diabetics for a patient with a better accuracy. This results its capable of predicting the diabetic effectively. We have studied the existence and outcomes using conventional risk factordiabetics

Reference	Objective	Method used	Result	Advantages	Disadvantages
[5]	Author uses diabetes predictio n in the proposed system with various classifications such as Naïve Bayes, SVM, and Decision Tree.	Machine learning , Ad boost.	Ad boost decisio n suppor t system has a precisi on of 80.72 %.	General database uses 768 instances and 9 attribute s as local database for validatio n.	It is in point of fact to predict more in female women.
[7]	Author uses Six machine learning techniqu es which is a tool for diabetic’ s predictio n.	Machine learning , Decisio n Tree, SVM, Naive Bayes.	DT76 %, SVM-79.68 %, NB-78.01 %.	This method relates the conduct of various machine learning techniqu es and compute s which algorith m is superior.	One of the limitatio ns of this method is that performs some of the input variables were mislaid from the dataset used.
[1]	The current techniqu e utilizes supervis ed machine learning algorith m.	Support vector machine , Random Forest, Artificia l Neural Network Naive Bayes.	SVM-84.09 %, RF-87.07 %.	They are able to enhance diabetic’ s predictiv e conduct by using min max scaling process.	The detectio n precisio n of ANN and NB was Poorer.
[4]	To predict the classification model which is based on supervis ed ML and DL Techniques.	Least square support vector machine , Random Forest.	LSSV M-97.08 %, SVM-97.07 %, RF-88%	This Predictiv e analysis gives a high promine nce in the emergin g big data technology. High efficienc y and	Extracti ng Significant features for predictin g diabetics is quite complicated.

Table 1. Summary of ML Techniques on Diabetes Mellitus

III. DATA AVAILABILITY

The goal of the proposed work to analyze the diabetes dataset over the classification techniques. Our research concentrates to reduce the complications of diabetes through early prediction of diabetes and to improve the prognosis(lives) of the people. A person with diabetes has considerable features for the cause of disease which is depending on the age, glucose level, heredity, and other factors, as well as features vary from one type to another type. From the analyzed articles the mostly used dataset which is collected from UCI repository archives.ics.uci.edu-Diabetes. *We have a sample diabetic dataset here, comprising of 15 attributes, and its description of attributes is given Training and Testing of data over the classification techniques, we have considered 768 data items.*



S.NO	ATTRIBUTE	DESCRIPTION
1	Age	Age of a person
2	Gender	Male or female
3	Plasma glucose fasting	-
4	Plasma glucose post prandial	-
5	Pregnancy	Pregnancy count of women
6	Blood Glucose Level	Plasma glucose concentration a 2h in an oral glucose tolerance test
7	Blood pressure	Diastolic blood pressure (mm Hg)
8	Skin thickness	Triceps skin fold thickness(mm)
9	Insulin	2-h serum insulin (mu U/ml)
10	BMI (Body mass index)	Body mass index (weight in kg/(height in m)^2)
11	DPF	Diabetes pedigree function
12	Serum creatinine	Test measures the level of creatinine in the blood
13	Serum sodium	Sodium content in your blood
14	Serum Potassium	Potassium content in blood
15	HBA1C	Hemoglobin A1c, a blood pigment that carries oxygen

Table 2.Description of Attributes

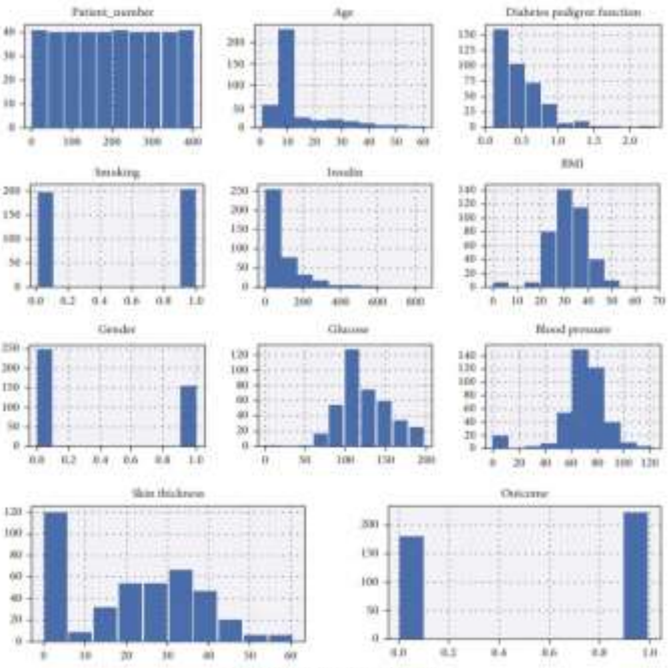
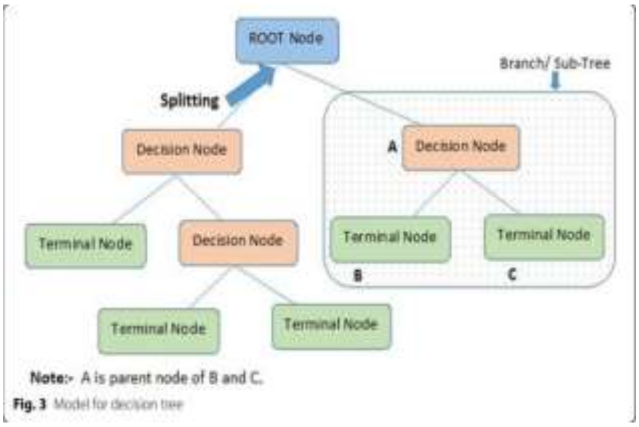


Figure 2: Histogram of each attribute.

IV. VARIATIONS OF MACHINE LEARNING ALGORITHMS:

A.DECISION TREE

Decision Tree is a supervised learning method, which is used for resolve classification problems. The goal of the method is forecast the class value of get target variable. This gives decision tree a lead of choosing the most compatible hypothesis among the training dataset. Input**:** training data set **Output:** decision model (tree structure) [2].

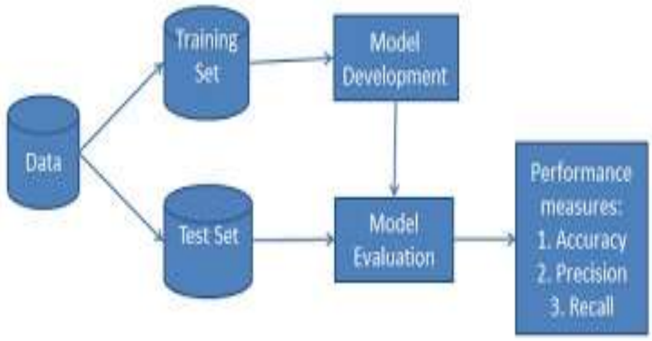


Note:- A is parent node of B and C.

Fig. 3 Model for decision tree

B.NAIVE BAYES

Naive Bayesian method takes and improvised the dataset as input, performs analysis and implies the class label using Bayes Theorem. It computes a probability of class in input data and provides to predict the class of the unused data sample [2].



Formula:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

•p (c | x) is the eventual probability of class (target) given predictor (attribute).

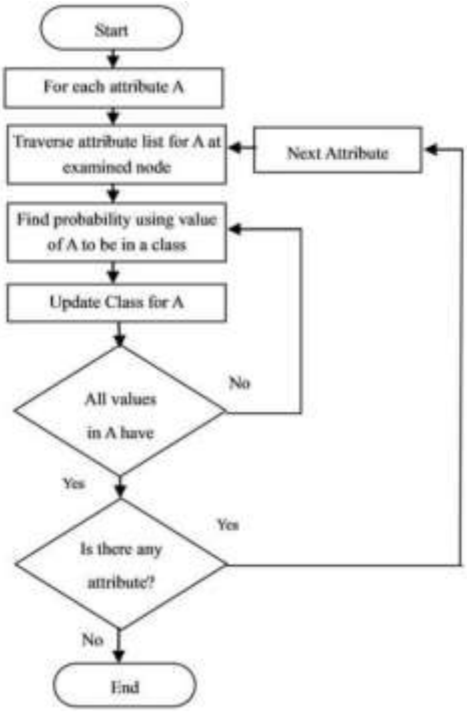
• P(c) is the superior probability of class.

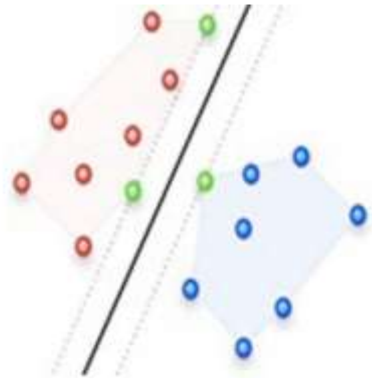
•P (x | c) is the liability which is the probability of forecasting given class.

•P(x) is the leading probability which forecast Support vector machine.

C.SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised learning, differential classification technique. This technique can be used for both regression and classification. The SVM training algorithm constructs a model that conserves new samples to one of the classes [2]. SVM can distinguish both continuous and discrete data as it automatically normalizes the data before they are modelled. It was actually developed for solving the binary classification problems. Its usability has now extended to make it suitable to support multi class data and regression problems. However, primarily it is used for classification problems in machine learning. SVM can also be used in the KNN classifier. It becomes difficult to imagine when the number of features exceeds more. SVM have their unique way of implementation as compared to other machine Learning Algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.





D.RANDOM FOREST

Random Forest is a supervised learning which is used for both classification and Regression. The random forest is the proceeding process of ruling the root node and separating the component node will run at any rate. The Steps given below are:

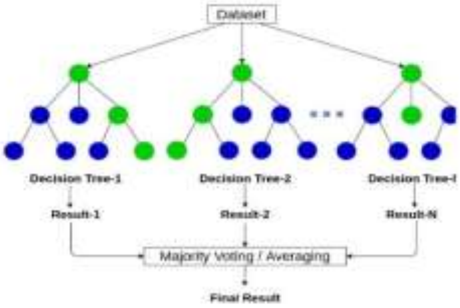
Load the data where it consists of “m” features emphasis characterizing the attribute of the dataset.

The training algorithm of random forest is also called as bootstrap algorithm or extending technique to select n component at any rate from m features(i.e.) to create arbitrary samples, this model trains the new sample to out of bag sample (1/3rd of the data) used to decide the impartial OOB error.

Compute the node d using the best split. Divide the node into sub-nodes.

Repeat the steps, until n number of trees.

Compute the total number of votes of each tree for the forecasting target. The highest majority class is the final projection of the random forest [2].



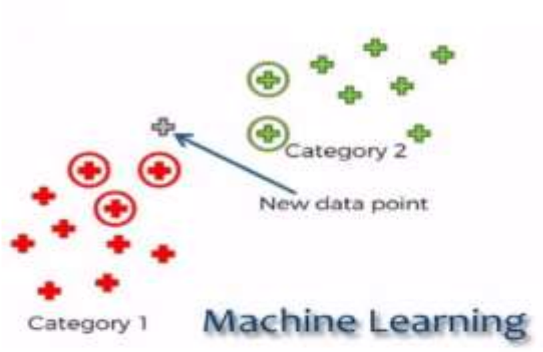
E.K NEAREST NEIGHBOUR (KNN)

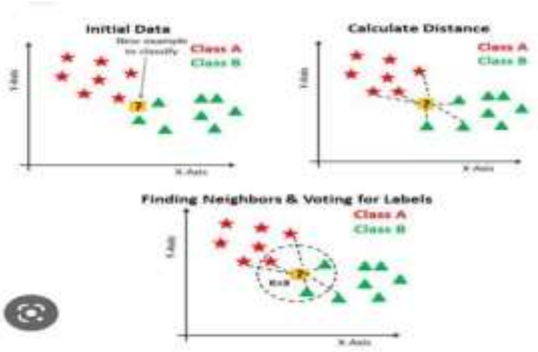
KNN is a classification method which classifies the new sample based on closeness measure or distance measure. The steps for KNN are given below:

Training aspect of the algorithm consists of only conserving the feature sample and class label of training sample.

Classification aspect: the user has to describe a "k" value for the classification of the enduring sample for the k number of the class labels, so the unlikable sample can be classified into the determined class based on the feature comparison.

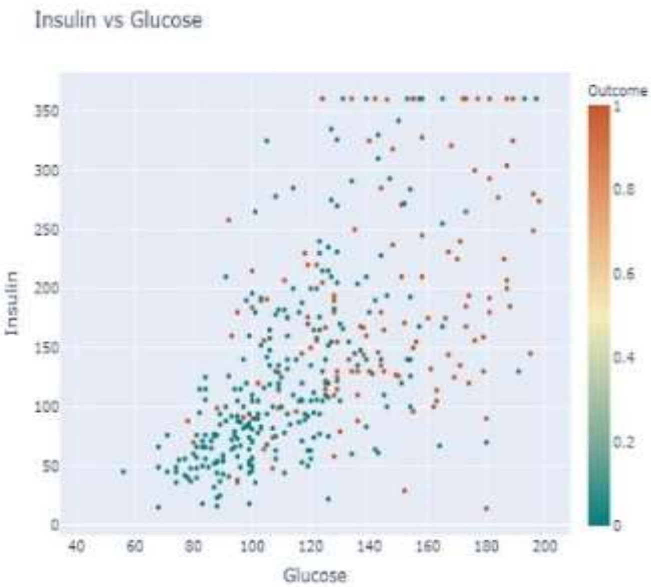
Mass of voting classification occurs for unlikable class. The value of the k can be selected by various takings like heuristic method [2]. It is represented by the diagram below:

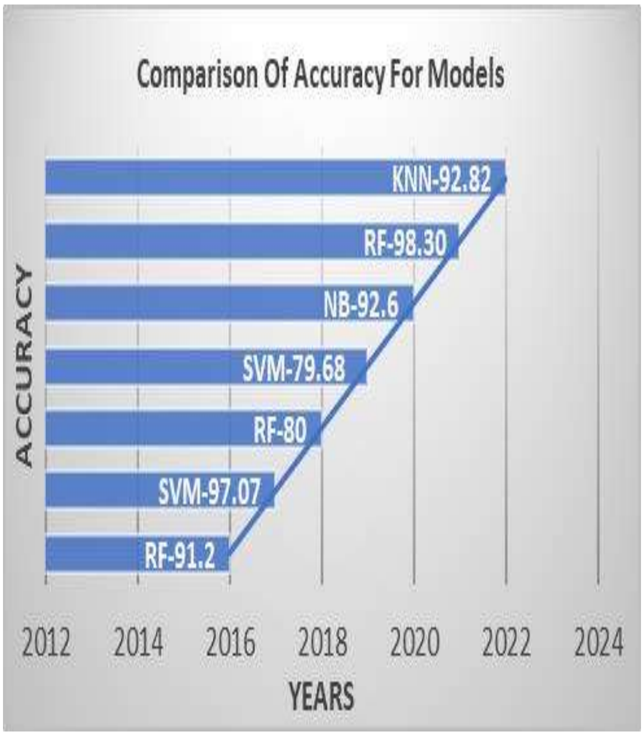




V. RESULTS AND DISCUSSIONS

With respect to applications of machine learning and data mining methods the present study was reviewed in diabetic research. Thus, each article was accordingly categorized and analyzed article comparison was carried out based on the previous retrieved articles. The above techniques are analyzed which gives an insight of various MLAlgorithms. In contrast to other classification models Random Forest, Naïve bayes, Support vector machine (SVM) which results and performs the better accuracy of overall analyzed papers. The top performing was Random Forest which it generates early-stage prediction of diabetic thus it is acknowledged that they are not statistically significant with diabetes. According to some surveys which have been published. The authors described only the studies related to Decision Tree, Support Vector Machine, Artificial Neural Network and some DL techniques. It also surveyed the main well-known classification techniques to predict diabetes.





VI. CONCLUSION AND FUTURE SCOPE

In this paper we come to analysis that they have trained various ensemble models mostly using Pima and Clinical dataset. The Correlation based feature method improvised the performance of the Model to determine the best and most diabetes Prediction Algorithm, a variety of various Algorithms and combinations of algorithms can be examined by the existing analysed Models. We can use Artificial intelligence, Deep Learning and Reinforcement learning for future enhancements to predict diabetes of larger datasets to examine higher accuracy. In the analysed articles LGBM is a higher accuracy at a maximum were compared to a RF and GB classifiers. This Overview helps to provide a clear-cut view of diabetes prediction and helps to frame better. Diabetes Prediction techniques to overcome diabetes through timely prediction. More Over we have analysed and evaluated different Schemas for Optimal performance and results. This may be capable to predict the chances of degrees of diabetes and gives the first-class getting to know set of rules with better accuracy comparatively. The core objective of future is to enhance the accuracy of predictive model This accuracy can be increase by improving the performance of the data, the algorithms or even by algorithm tuning. This would be accomplished by gathering diabetic patient's datasets from various sources, to generate a better relevant prototype. This is a limitation of this research.

REFERENCES

[1] "Learning about diabetes and its type." [Online]. Available: https://www.diabetesresearch.org/what-is-diabetes

[2] D. Gahlan, R. Rajput, and V. Singh, "Metabolic syndrome in north indian type 2 diabetes mellitus patients: A comparison of four different diagnostic criteria of metabolic syndrome." Diabetes & metabolic syndrome, vol. 13, no. 1, pp. 356–362, 2018.

[3] "Causes of diabetes that must be known." [Online]. Available: https://markethealthbeauty.com/causes-of-diabetes

[4] "Learning of diabetes facts figures." [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[5] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: a systematic review," IEEE Journal of Biomedical and Health Informatics, 2020.

[6] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis, vol. 992. Springer, Singapore, 2020, pp. 113–125.

[7] L. Li, "Diagnosis of diabetes using a weight-adjusted voting approach," in 2014 IEEE International Conference on Bioinformatics and Bioengineering, 2014, pp. 320–324.

[8] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," Procedia Computer Science, vol. 47, pp. 45–51, 2015.

[9] P. Agrawal and A. kumar Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus." International Research Journal of Engineering and Technology, vol. 02(03), 2015.

[10] N. Naiarun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," Procedia Computer Science, vol. 69, pp. 132– 142, 2015, the 7th International Conference on Advances in Information Technology.

[11] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017. [12] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International Journal of Medical Informatics, vol. 97, 2016.

[13] T. Ahmed, "Developing a predicted model for diabetes type 2 treatment plans by using data mining," Journal of Theoretical and Applied Information Technology, vol. 90, pp. 181–187, 2016.

[14] A. Oleiwi, L. Shi, Y. Tao, and L. Wei, "A comparative analysis and risk prediction of diabetes at early stage using machine learning approach," International Journal of Future Generation Communication and Networking, pp. 4151–4163, 2020.

[15] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," Complex., vol. 2021, pp. 5 525 271:1–5 525 271:10, 2021.

[16] "Pima. university of california, irvine, school of information and computer sciences," Oct 6, 2020. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[17] T. Chen and C. Guestrin, "Xgboost: Reliable large-scale tree boosting system," in Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2015, pp. 13–17.

[18] E. Goel, E. Abhilasha, E. Goel, and E. Abhilasha, "Random forest: A review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 1, 2017.

[19] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275–2