WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Diabetes Prediction using RF based classifier using machine learning:- A Review

[1]Gudla Nikhila

[1]Student, Department of Information Technology, G M R Institute of Technology, Rajam, A.P
20341A1231@gmrit.edu.in

ABSTRACT:

One of the most quickly spreading diseases, diabetes (also known as diabetes mellitus) has impacted millions of people worldwide. A chronic condition known as diabetes is characterized by elevated blood sugar. It may result in a variety of complex diseases, including heart attack, kidney failure, and stroke. Prediction, appropriate treatment, and management are all crucial. Understanding the disease's symptoms is crucial for making predictions about it. Currently, diabetes prediction also uses machine-learning techniques. Diabetes can be predicted using machine learning methods including logistic regression, decision trees, random forests, and naive bayes. Compared to other algorithms, random forest predicts diabetes more accurately in this case. Numerous decision trees make up the categorization technique known as random forest. When constructing each individual tree, it employs bagging and feature randomness in an effort to produce an uncorrelated forest of trees whose forecast by committee is more accurate than that of any individual tree. certain datasets. This is one method that machine learning can be used to anticipate diabetes.

Keywords: -Diabetes Mellitus, Logistic regression, Decision Tree, Random Forest, Naïve Bayes.

INTRODUCTION

Diabetes is a category of metabolic illnesses characterized by persistently elevated blood sugar levels. Frequent urination, increased thirst, and increased appetite are common symptoms. Numerous health issues might be brought on by diabetes. Cardiovascular illness, stroke, chronic kidney disease, foot ulcers, and cognitive impairment are examples of serious long-term consequences. Diabetes results from either insufficient insulin production by the pancreas or improper insulin utilization by the body's cells. A hormone called insulin is in charge of facilitating the entry of food-derived glucose into cells for cellular energy utilization. Diabetes can come in three different forms.

1. The loss of beta cells in the pancreas causes type 1 diabetes, which is caused by an inability to produce enough insulin. Previously, this type was referred to as "juvenile diabetes" or "insulin-dependent diabetes mellitus."
2. Insulin resistance, a disease in which cells do not react to insulin as intended, is the precursor of type 2 diabetes. A shortage of insulin may also develop as the condition worsens. Previously, this type was referred to as "adult-onset diabetes" or "non-insulin dependent diabetes mellitus.
3. "Gestational diabetes is the third major type, and it happens to pregnant women who have never had diabetes before because of elevated blood sugar levels. After delivery, blood sugar levels in women with gestational diabetes typically return to normal.

Because the majority of medical data are nonlinear, non-normal, correlation-structured, and complicated in nature, analyzing diabetic data can be difficult. Furthermore, feature selection techniques (FST) and classifiers can also be employed with ML-based systems. Additionally, it aids in the proper diagnosis of diabetes, with the best classifier serving as the key to determining an individual's risk for developing the disease. Different machine learning (ML)-based systems, such as naive Bayes (NB), support vector machine (SVM), Adaboost (AB), decision tree (DT), and random forest, were employed to categorize and predict diabetes illness (RF). After data analysis, machine learning approaches aid in the early detection and prediction of diabetes. This study examines the effectiveness of random forest ML algorithms for early diabetes prediction.

LITERATURE SURVEY

In the paper[1] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers,"

- In this Paper,Author proposes a new pipeline for diabetes prediction from the PIMA Indians Diabetes dataset.

- Author used XG boost, Naïve Bayes,Logistic Regression, Random Forest,Decision Tree and Adaboost.
- Finally, Author concluded that Random Tree-Based classifiers are well suited for the data to be classified when inter-class redundancy is much higher as in the PID dataset.

In the paper [2]Maniruzzaman, M., Rahman, M., Ahammed, B., & Abedin, M. (2020). Classification and prediction of diabetes disease using machine learning paradigm.

- *The main objective of author study is to identify the most significant factors of diabetes disease based on LR model and develop a ML-based system for the accurate risk stratification of diabetes disease*
- *He used Naïve Bayes,combination of Logistic Regression &Random Forest, Decision Tree and Adaboost.*
- *He used in ML based system using LR-RF combination for feature selection technique and classifier gave the highest classification accuracy.Finally, he got the accuracy of 94.25%.*

In the paper [3] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data with Missing Values,"

- *In this paper, they have proposed a diabetes mellitus classification algorithm targeting the challenging characteristics of missing values and class imbalance problems seen in the diabetes dataset. The algorithm has addressed both the data preprocessing and classification phases.*
- *Author uses Naïve Bayes, Adaptive synthetic sampling method (ADASYN), Random Forest Algorithms.*

In the paper [4] T. M. Le, T. M. Vo, T. N. Pham and S. V. T. Dao, "A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic,"

- *This research performs an Artificial Neural Network (ANN) method with a metaheuristics-based feature selection algorithm to enhance performance.*
- *They compare results with several conventional machine learning algorithms such as Support Vector Machine, Decision Tree, Random Forest Classification, Naive Bayes Classification, Logistic Regression.*
- *They got better accuracy in Random Forest.*

In the paper [5] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab and S. A. C. Bukhari, "Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review,"

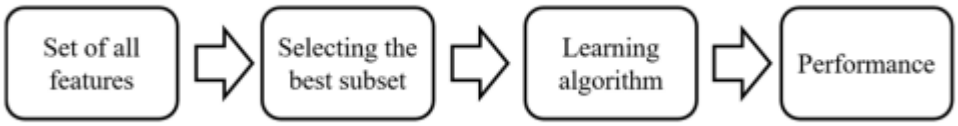
- *Firstly, Author explores and investigate the data mining based diagnosis and prediction solutions in the field of glycemic control for diabetes.*
- *Secondly, he provides a comprehensive classification and comparison of the techniques that have been frequently used for diagnosis and prediction of diabetes based on important key metrics*
- *Based on his analysis and evaluation, they conclude that for accurate detection, classification, and prediction of the disease, they need to preprocess the data and use hybrid techniques.*

METHODOLOGY

- This is the classification model in which the goal is to predict the discrete value like {0,1} or (yes, no) or (spam, not spam). Here we are predicting whether the person is having diabetes or not as like as (yes or no).
- There are three types of classifications:
    1. Binary Classification: Classification task with two possible outcomes.
    2. Multi-class Classification: Classification with more than two classes.

    3. Multi-label Classification: Classification task where each sample ismapped to set of targetlabels (more than one class).
- Methods used for classification are:

1. Logistic Regression
2. Naïve Bayes
3. Decision Tree
4. Random Forest
5. Support Vector Machine
6. Ada Boost

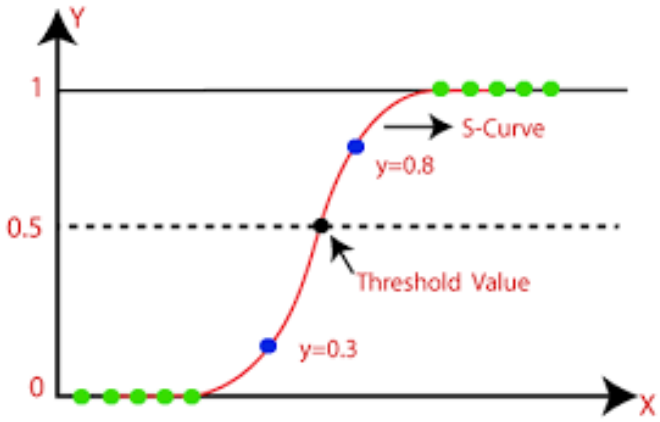


Logistic Regression:

Definition: A machine learning algorithm for classification is logistic regression. In this approach, a logistic function is used to model the probabilities describing the potential outcomes of a single experiment. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1.

Advantages: For this aim (classification), logistic regression was created, and it works best when examining the impact of several independent variables on a single outcome variable.

Disadvantages: It only works when the predicted variable is binary, all predictors are independent of each other, and the data is free of missing values.

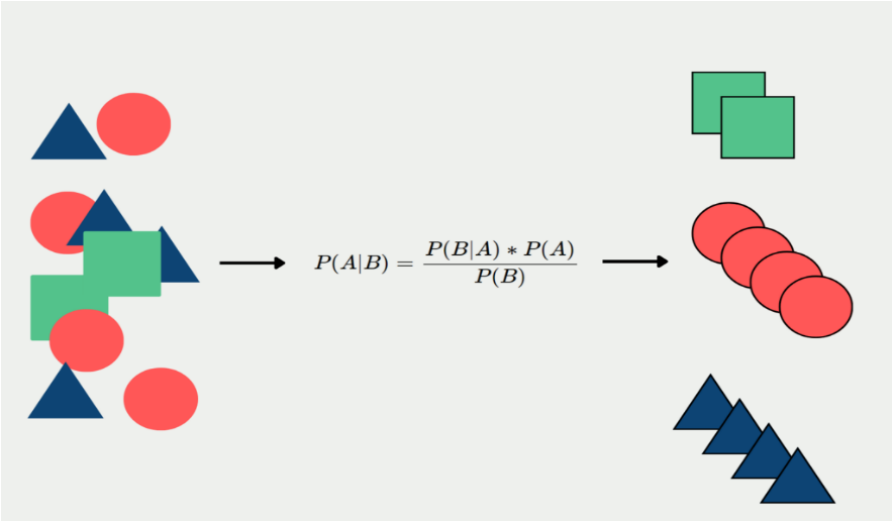


Naive Bayes:

Definition: The Naive Bayes algorithm is based on the Bayes theorem and assumes independence between every pair of features. Naive Bayes classifiers are a type of classification algorithm that is based on the Bayes Theorem. It is a family of algorithms that all share a common principle, namely that every pair of features being classified is independent of each other. Naive Bayes classifiers perform well in a wide range of real-world applications, including document classification and spam filtering.

Advantages: To estimate the required parameters, this algorithm requires only a small amount of training data. When compared to more sophisticated methods, Naive Bayes classifiers are extremely fast.

Disadvantages: It's well known that Naive Bayes is a poor estimator.

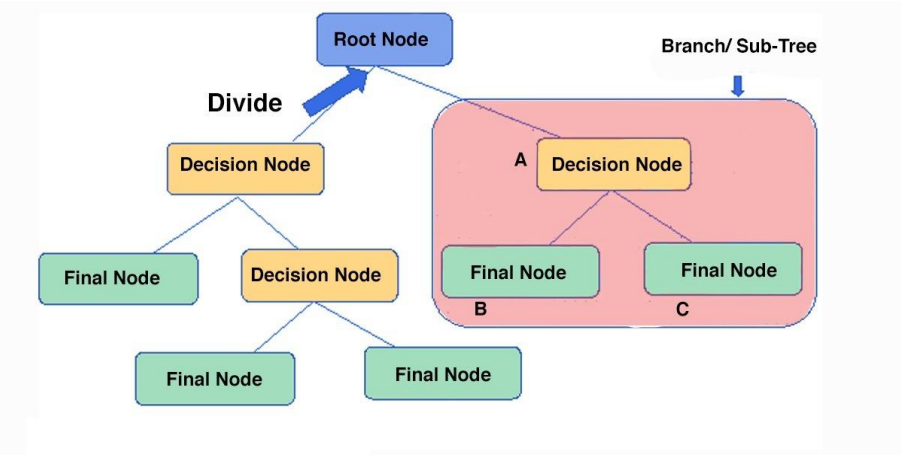


Decision Tree:

Definition: A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. A decision tree generates a set of rules that can be used to categorise the data given a set of attributes and their classes.

Advantages: Decision trees can handle categorical and numerical data, are easy to understand and visualize, and require little data preparation.

Disadvantages: Decision trees can build complicated trees that do not generalize well, and they can also be unstable since little changes in the data may lead to the generation of a completely different tree.

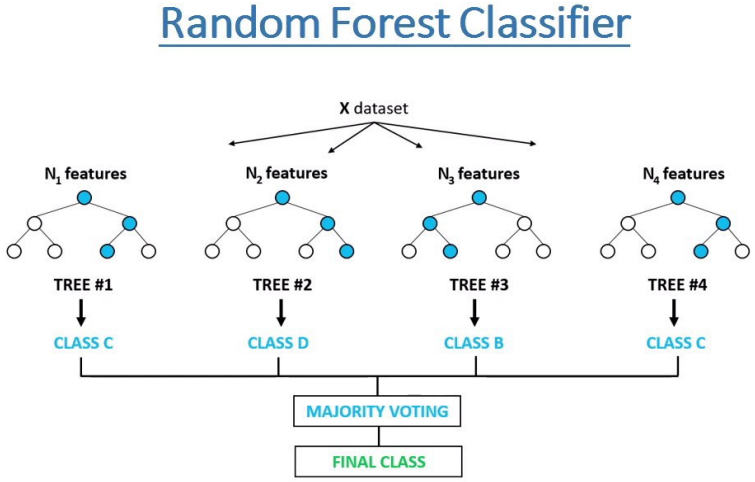


Random Forest:

Definition: Random Forest classifier is a meta-estimator that fits several decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Advantages: Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.

Disadvantages: Slow real time prediction, difficult to implement, and complex algorithm.

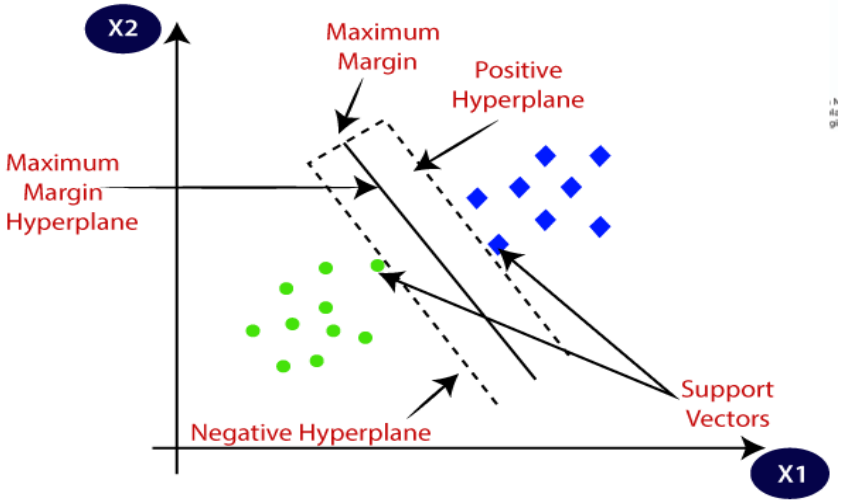


Support Vector Machine:

Definition: A support vector machine represents the training data as a set of points in space divided into groups by a distinct gap that is as large as possible. Then, based on which side of the gap they fall, new samples are projected into that same area and predicted to belong to a category. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary.

Advantages: When there is a clear margin of dissociation between classes, support vector machines perform comparably well. In high dimensional spaces, it is more effective.

Disadvantages: This algorithm does not operate on datasets with more noise and is not appropriate for huge data.

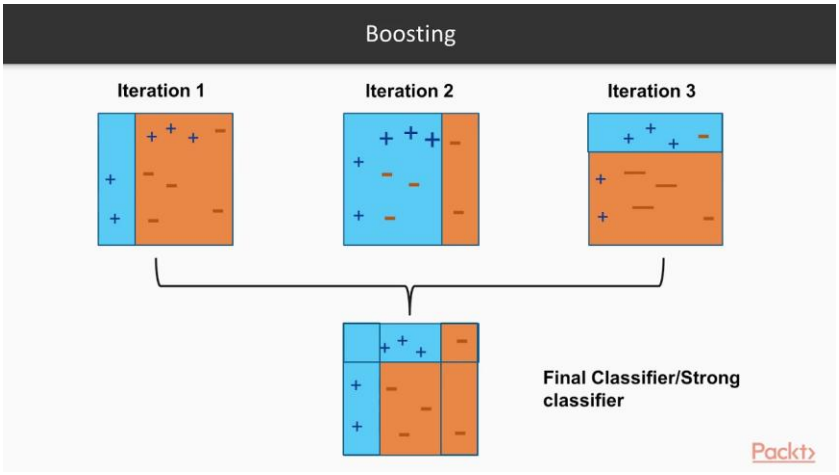


Ada Boost:

Definition: A meta-estimator called an AdaBoost classifier starts by fitting a classifier to the initial dataset. It then fits additional copies of the classifier to the same dataset, but with the weights of instances that were mistakenly categorized being changed such that later classifiers would concentrate more on challenging cases. AdaBoost, also known as Adaptive Boosting, is an ensemble method used in machine learning. Decision trees with only one split, or those with one level, are the most frequent algorithm employed with AdaBoost. Decision Stumps is another name for these trees.

Advantages: Since the input parameters are not simultaneously optimized, Adaboost is less prone to overfitting. By applying Adaboost, weak classifiers' accuracy can be increased.

Disadvantages: A good dataset is necessary for Adaboost. Before implementing an Adaboost algorithm, noisy data and outliers must be avoided.



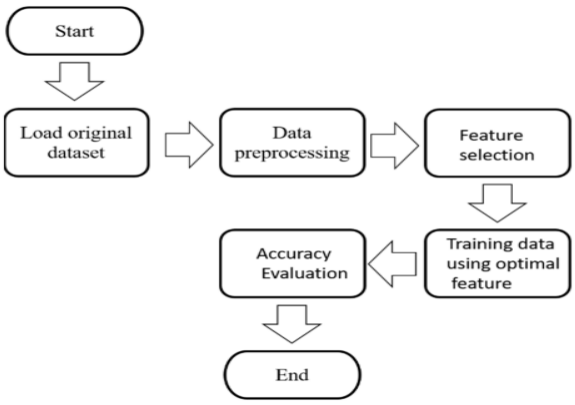
4.RESULTS AND DISCUSSION

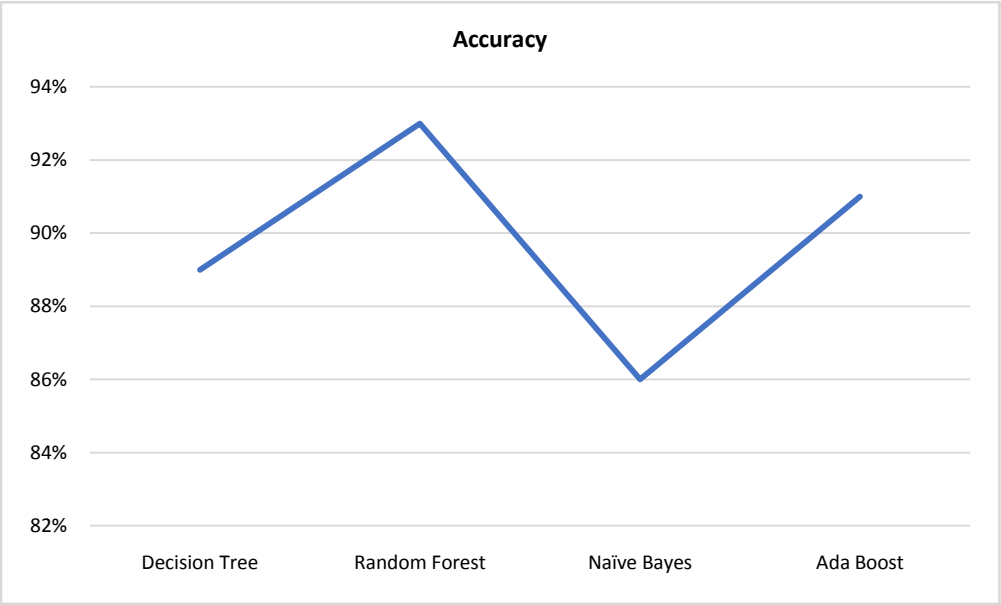
- In this study, the ML-based system to classify the patients into two classes as diabetic or not a diabetic.
- By considering research papers the outputs are given below.

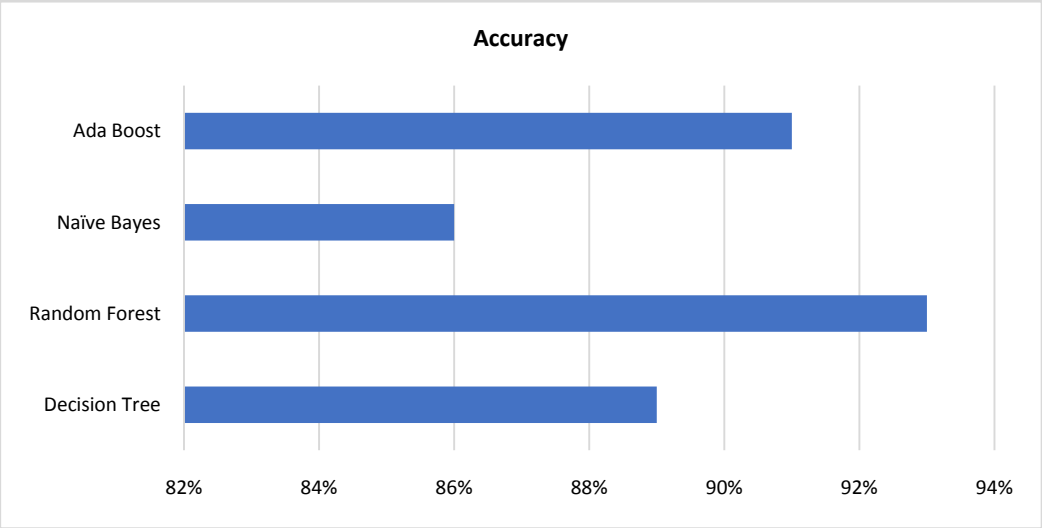
S. No	Method	Accuracy
1.	Decision Tree	89%
2.	Random Forest	93%
3.	Naïve Bayes	86%
4.	Ada Boost	91%

In this paper, we have reported the risk stratification using an ML-based method to divide the patients into diabetic and control groups. We used a diabetes dataset from the National Health and Nutrition Examination Survey that was collected between 2009 and 2012. 6561 respondents make up the sample, including 657 diabetics and 5904 controls. Decision tree classifier accuracy is 89%, Random Forest classifier accuracy is 93%, Naive Bayes classifier accuracy is 86%, and Ada Boost classifier accuracy is 91%.

We achieved the highest accuracy for Random Forest by considering the accuracy.







5.CONCLUSION

From the analysis conducted in this paper we have attempted to explain and compare the performances of different machine learning algorithms on diabetes prediction. Specifically we identified a trends with respect to the types of machine learning methods being used, the types of training data being integrated. The various methods applied are Ada Boost, Naïve Bayes, Decision Tree and Random Forest. Among all Random Forest gives the more accuracy in this diabetes prediction. We can increase the accuracy by cross validation.

REFERENCES

[1] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in *IEEE Access*, vol. 8, pp. 76516-76531, 2020.

[2] Maniruzzaman, M., Rahman, M., Ahammed, B., & Abedin, M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health information science andsystems.

[3] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on

Imbalanced Data With Missing Values," in *IEEE Access*, vol. 7, pp. 102232-102238, 2019.

[4] T. M. Le, T. M. Vo, T. N. Pham and S. V. T. Dao, "A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic," in IEEE Access, vol. 9, pp. 7869-7884, 2021.

[5] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab and S. A. C. Bukhari, "Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review," in IEEE Access, vol. 9, pp. 43711-43735, 2021.

[6] U. Ahmed et al., "Prediction of Diabetes Empowered with Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022.

[7] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," in IEEE Access, vol. 7, pp. 144777-144789, 2019.

[8] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis and K. Moustakas, "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction," in IEEE Access, vol. 9, pp. 103737-103757, 2021.

[9]M. S. Islam, M. K. Qaraqe, S. B. Belhaouari and M. A. Abdul-Ghani, "Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes," in IEEE Access, vol. 8, pp. 120537-120547, 2020.

[10] Mohapatra SK, Swain JK, Mohanty MN. Detection of diabetes using multilayer perceptron. In: International conference on intelligent computing and applications, 2019, pp. 109–116.

[11] Pei D, Zhang C, Quan Y, Guo Q. Identification of potential type II diabetes in a Chinese population with a sensitive decision tree approach. J Diabetes Res. 2019; 2019:1–7.

[12] S. Perveen, M. Shahbaz, T. Saba, K. Keshavjee, A. Rehman and A. Guergachi, "Handling Irregularly Sampled Longitudinal Data and Prognostic Modeling of Diabetes Using Machine Learning Technique," in IEEE Access, vol. 8, pp. 21875-21885, 2020.

[13] R. Marzouk, A. S. Alluhaidan and S. A. El_Rahman, "An Analytical Predictive Models and Secure Web-Based Personalized Diabetes Monitoring System," in IEEE Access, vol. 10, pp. 105657-105673, 2022.

[14] H. Kaur and V. Kumari, ‘‘Predictive modelling and analytics for diabetes using a machine learning approach,’’ Appl. Comput. Informat., Dec. 2018.

[15] H. Tong, B. Liu, and S. Wang, ‘‘Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning,’’ Inf. Softw. Technol., vol. 96, pp. 94–111, Apr. 2018.