Early Diagnosis of Diabetes Using an Ensemble Classifier for an Optimal Number of Features

Rajasekhar C[1], Bibin Varghese[2], Dr. Smita C Thomas[3]

[1,2,3]Computer Science and Engineering, Mount Zion College of Engineering, Kadammanitta Pathanamthitta, Kerala, India

---***---

Abstract - *In the healthcare industry, data mining is essential for disease prediction. In the early diagnosis of disease, data mining techniques are commonly applied. Diabetes is one of the world's most serious health issues. A widespread chronic condition is a diabetes. Diabetes prediction is a science that is increasingly growing. Diabetes prediction at an early stage will lead to better therapy. It is necessary to avoid, monitor, and increase diabetes consciousness because it causes other health issues. Diabetes of type 1 or type 2 can lead to heart disorders, kidney diseases, or complications with the eye. By using different approaches, Diabetes can be forecasted from the patient's dataset. In this study, significant features are used to predict diabetes, and the link between the various traits is also described. This paper focuses on the result-oriented analysis of the PIMA diabetes dataset for predicting diabetes among patients. RFE algorithm is used as a feature selection technique that has detected 5 features for the optimal working of the model.*

Key Words: Data Mining, Diabetes, Association Rule Mining, Feature Selection, Recursive Feature Elimination, Random Forest, XGboost Classifier.

1.INTRODUCTION

Data mining (DM) is used to invent and view information from data in a readily interpreted condition for humans. It is a method for reviewing vast quantities of data. The application of data mining strategies in different industries such as finance, education, and so on plays a significant role in IT [1]. DM can be successfully extended to disease prediction using different data mining approaches in the area of the medical domain. DM appears to be expected and defined in two dominant objectives. The prediction uses several variables or fields within the data set to forecast such variables of importance, either uncertain or potential values. The focus of the explanation is on the trends of human perception. Before designing predictive models, it is very important to understand the development and characteristics of impacting diabetes from external sources. The idea is to foresee diabetes & to classify diabetes causes by using data extraction methods [2].

The word diabetes is a condition occurring when the blood glucose in the body is too high and is also known as blood sugar. As described above, diabetes can lead to other significant cardiovascular complications. According to the WHO, there are 3.7 million deaths by age 70 years and cardiovascular disease is responsible for this high death rate. The key cause behind diabetes is uncontrolled blood glucose levels. A chronic condition is a continuous disease or condition whose consequences are permanent. This illness forms have a significantly detrimental effect on the quality of life. Diabetes is one of the most serious and widespread diseases. This chronic illness is the leading cause of death in adults globally. Chronic diseases are also linked to higher costs. Governments and people expend a large part of the budget on chronic diseases [3,4]. Worldwide diabetes statistics showed that about 382 million people worldwide suffered from this disease [5].

Diabetes is a long-term disease that affects the human body by reducing the amount of glucose-carrying insulin in the blood cells. This raises the level of sugar in the body which leads to multiple complications, including stroke, cardiac disease, blindness, kidney failure, and death [6]. Diabetes can be a chronic sickness with no illustrious cure if it insists on maintaining a traditional level of blood glucose without inducing hypoglycemia. Diet, exercise, and the use of relevant medicinal products are manageable. Diabetes Mellitus takes place around the world, and in developing worlds it's a big deal. The rise in rates in developed countries is a result of the movement towards rapid urbanization including changes in lifestyles, such as a "West" diet. This is attributed to a lack of understanding. Diabetes diagnosis for quantitative analysis is a complex problem. Certain parameters such as A1c, fructosamine, the number of white blood cells, fibrinogen, etc., are inefficient due to small levels. [7]. Diabetes is predicted in our study with the help of significant features and the correlation of the various attributes. Random feature elimination, random forest classifier, and XGBoost classifier were used to investigate diabetes diagnosis.

This is the rest of the paper: The related work in diabetes prediction is discussed in Section II. We said about a detailed methodology in Section III. Section IV tests and outcomes, while Section V concludes the article with future instructions.

2. LITERATURE SURVEY

Some of the research work which guides me in the way of completing my research paper is discussed below:

R. Syed (2018) This research proposes a classification technique that combines tree-based partitioning with an adaptive SVM approach. Data reduction was achieved in the suggested architecture by pre-processing under-sampling SMORT. The method is compared to traditional tree-based RF, RT, and J48 algorithms and tested utilizing the Weka tool on a diabetic dataset. The results show that the suggested algorithm outperforms the present method for analyzing diabetic data and producing an effective categorization. [8]

Z. Xu and Z. Wang (2019) Using the ensemble learning method, construct a type 2 diabetes risk prediction model. The suggested prototype uses a weighted feature selection technique dependent on random forest (RF-WFS) and the extreme gradient boosting (XGBoost) classifier for optimal feature selection. By comparing numerous performance indicators and the outcomes of various contrast studies, the method's efficiency was confirmed. Furthermore, the method outperforms other classification algorithms in terms of prediction accuracy (Random Fores (RF)t, C4.5, Naive Bayes (NB), AdaBoost). The validation findings for UCI Pima Indian diabetes dataset show that the prototype is more precise and better than prior literature study outcomes. As a consequence, it was shown that the prototype can be utilized for early step diagnosis of diabetes. [9].

R. S. Raj et al. (2019) examine electronic health information of diabetic patients from several sources to offer a medical case in this study. NB and SVM are two DM categorization methods utilized in this research. The goal of the study would be to use a health record to predict diabetes and compare the accuracy of these two algorithms to find a better diabetes prediction algorithm. [10].

S. A. Aboalnaser et al. (2019) A database about this disease was discussed in this research, and DM methods were used to implement it. Data mining methods are utilized to assist in predicting DM. It creates the prediction process quicker, cheaper, and more precise for the benefit of both doctors and patients. Well-known data mining strategies are explored in this research to achieve DM prediction. Using the Orange DM tool, the performance of various methods was assessed and discussed. The recall and precision measures were utilized to assess the efficiency of every categorization algorithms mentioned. The categorization approaches investigated were NB, Artificial Neural Network (ANN), Decision Tree, K-Nearest Neighbors, RF, SVM, and Logistic Regression. [11]

R. Karthikeyan et al. (2019) This work introduces the Rule-Based Classification (RBC) technique, which considers the best classifier to increase prediction accuracy in the medical data set. A diabetic data set is used in the proposed RBC method. The biggest drawback of diabetes is that not everyone experiences its symptoms, mandating diabetic testing. Rule-based systems are customizable and can be used to solve a variety of problems. By combining many stages, facts, and symptoms to build relevant rules and find the best rule for the condition, the RBC technique can be used to forecast diabetes in patients. This study also includes a comparison of different diabetic data set classifiers. [12].

Talha Mahboob Alam et al. (2019) [Base Paper] In this study, significant features are used to predict diabetes, and the link between the various traits is also described. To assess relevant attribute selection in diabetes, several techniques are used, As clustering, prediction, and ARM. The essential characteristics were determined using the principal components analysis technique. According to the Apriori method, there is a significant connection between diabetes, body mass index, and glucose levels. The use of ANN, RF, and clustering of K means was predicted to diabetes. [13].

M. Rout and A. Kaur (2020) Early prediction and healthcare diagnostic machine learning techniques must be more accurate, using accessible clinical database parameters and structures. The goal of this study is to look into and evaluate various machine learning approaches utilized in diabetic Mellitus, as well as the efficiencies gained, which could be used in the future to construct a predictive diabetes model. This research seeks to explore and investigate different results from the study of machine learning methods utilized in diabetic Mellitus and the efficiency of developing a future predictive diabetes model. [14]

A. P et al. (2021) In this study, the suggested research is primarily concerned with the most accurate diabetes prediction in patients. It is feasible to forecast the outcome or determine whether or not the patient is affected using classification algorithms based on diabetic data. A predictive model was constructed using five DM categorization algorithms, including NB, SMO, Multiclass, RF, and IBK, to predict early-stage diabetes and measurement precision. Each of the five methods is evaluated using different criteria, such as accuracy, recall, and precision [15].

3. PROPOSED METHODOLOGY

The artificial neural network was shown to be inefficient in predicting diabetes in the dataset in terms of accuracy and correctly and erroneously classified cases. As a result, a strategy for resolving this problem was proposed, which is discussed in this section: The first section explains how to collect diabetic data from standard UCI repositories; the second section explains how to select features using RFE and a random forest classifier; the third section explains hyperparameter selection, and the fourth section explains the voting classifier's classification procedure. And finally, the fifth section discusses the algorithm of the proposed work and flow diagram. Fig. 2 illustrates the workflow of the proposed work.

3.1 Data Collection

This study made use of the Pima Indian Diabetes (PID) Dataset, which was accessible through the UCI Machine Learning Repository.

3.2 Data Preprocessing

First, we performed some pre-processing on the raw data. It is possible to provide missing values and/or noisy and incorrect information. Pre-processing of data is necessary to ensure that high-quality results are produced. Data is pre-processed via the use of cleaning, integration, transformation, and discretization [16].

a) Data Cleaning

Data cleaning is used to eliminate duplicate values and distracting data. Outliers are required to overcome inconsistency in noisy data [17]. Several null (0) values, such as glucose, blood pressure, skin thickness, insulin, and BMI, are included in our data collection. As a result, the median value of all zero values was used as a substitute.

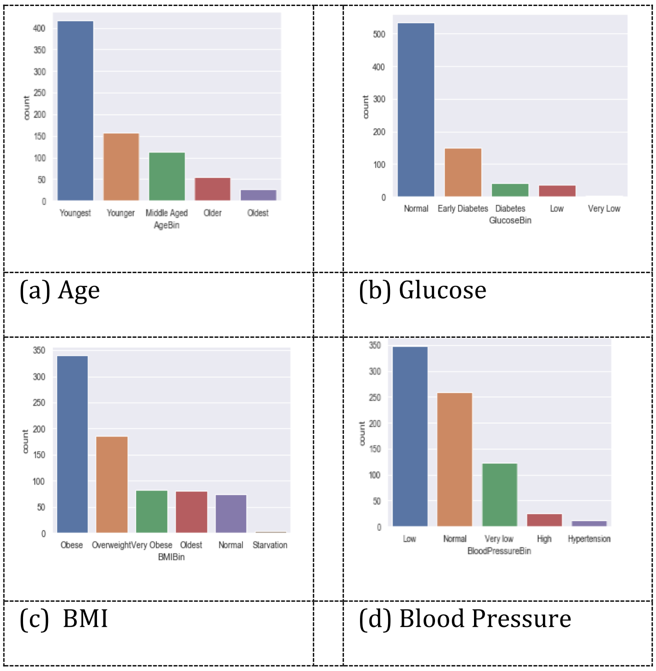
b) Data Reduction

Data reduction is the process of condensing a database into a smaller representation that gives nearly equivalent results. The process of lowering the number of features is referred to as "dimensionality reduction" [18]. To extract useful features from the entire database, the Recursive Feature Elimination using the RF method is applied. Glucose, BMI, diastolic BP, age, and insulin are all important aspects to consider.

c) Data Transformation

Smoothing, normalization, and data aggregation are the components of data transformation [19]. For scaling info, a regular scalar is being used here. The binning approach has been used to smooth the data. The age attribute is beneficial in five groups, as shown in figure 1(a). In patients without diabetes, the concentration of blood glucose is distinct from diabetes patients. As seen in Diagram 1(b), glucose levels were classified into 5 categories [20]. There is a significant connection between BMI & diabetes. Globally, there is a rise in the incidence of diabetes and heart disease. Furthermore, a previous study has revealed that BMI is a substantial risk factor for type 2 diabetes. [21]. As seen in diagram 1(c), BMI values were divided into 5 groups. There has been a direct correlation in their blood pressure levels between healthy or diabetic patients [22]. As seen in diagram 1(d), blood pressure was divided into 5 groups.

Selection of important attributes & translation of important attributes into bins was carried out after data cleaning to complete the pre-processing task.



(a) Age

(b) Glucose

(c) BMI

(d) Blood Pressure

Fig- 1: Graph of all binned features

Pre-processing, collection, and translating into bins are used to achieve critical qualities after data cleansing.

3.3 Feature Selection Using RFE

Feature choice is the process of making selections from several properties or characteristics such that computer latency and complexity are decreased and precision is improved and overfitting is prevented. [23].

a) Recursive Feature Elimination

The feature selection algorithm RFE, or Recursive Feature Elimination, is well-known. RFE is frequently used since it is easy to configure and use and efficient in identifying which characteristics (columns) are more or less important in defining the target variable in a training dataset. [24] In simple words, RFE ranks the features based on their importance and returns top-n features after eliminating the least important features, where n is given by the user.

Algorithm:

1. RFE uses a supervised learning estimator that has already been fitted to all features using data.

2. Then it considers the coefficient associated with each feature which it gets from the coef. or feature importance attribute. Basically, those coefficients are the same which we get after fitting the model on the dataset after minimizing the residuals. The value of these coefficients represents their importance with the target variable. The feature with the smallest absolute coefficient value is deemed the least significant, and so on.

3. The feature with the smallest absolute coefficient value is deemed the least significant, and so on. The number of features to be dropped at each iteration is taken from the step parameter. It is preferable to remove 1 feature at a time because the coefficient values of other features change when the model is rebuilt.

4. It rebuilds the model with each iteration, removing the least significant feature(s), and repeating the process until it only has two features. After that, it classifies features based on how long they took to be eliminated. The highest rank is given to the feature that was deleted first, and so on. Rank 1 is assigned to the last n features that have been deleted.[25]

3.4 HyperParameters Selection

Hyperparameters are the parameters that determine the representation architecture, and the process of hyperparameter tuning is used to find the optimal model architecture. These methods demonstrate how to use the space of probable hyperparameter values to represent the likely model structure. In our approach, we use Randomized-Search CV to adjust the voting classifier's parameters.

3.5 Association Rule Mining (ARM)

ARM is a significant part of DM. The techniques used by ARM [26] are widely used to discover hidden relations between objects in a transaction. For different businesses that help with various directive processes, mining association regulations from a large quantity of repository data are involved. A kind of mining specifies a careful pattern of locality that is simply calculated and interfaced. Support and confidence are the two major fundamental occurrences of the association rules. When two things are described as the proportion of occurrence of two objects & sums of all transactions, and the possibilities of taking into account the rules that arise from the circumstance, this transaction involves confidence.

a) Apriori

When another transaction database input is given, it is an ARM method that mines all frequently occurring objects in a transaction. If the PID data set is used as the Apriori input, this section generates a list of frequently occurring risk factors, indicating which variables contribute to the development of diabetes. Apriori is based on the concepts of trust and support. Regularly, the approach generates candidate itemsets for each n-itemset. Apriori may produce a large number of risk variables and these risk factors can be anticipated.

3.6 Voting Classifier

A Voting Classifier is an ML method that learns from a variety of samples and returns the maximum likelihood of the selected class.

This simply summarizes the voting classifier's results and forecasts the performance category with the greatest majority. We design a prototype that trains and forecasts efficiency dependent on a majority vote of the performance groups, rather than creating separate models for each of them. XGBoost and Random forest classifier and basic statistics (such as the average) are introduced in our proposed projects [27].

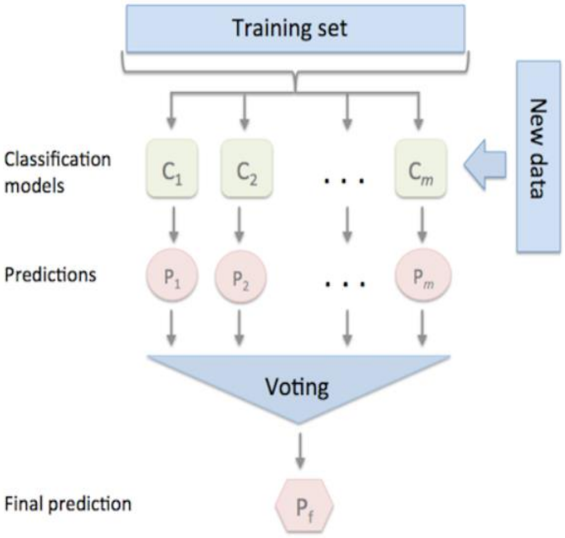


Fig -2: Voting Classifier

```
1.  Insert the data
2.  Insert model_selection
3.  Insert Bagging
4.  Insert Gradient Boosting
5.  ensemble insert VotingClassifier
6.  review=dataset. load review()
7.  x,y=review.data
8.  clf1=Bagging(random_state=1)
9.  clf2=Gradient
    BoostingClassifier(random_state=1)
10. eclf=EnsembleVoteClassifier(clfs=[clf1,clf2],weigh
    ts=[1,1])
11. labels = ['XGBoost, 'Random Forest Classifier',
    'Ensemble']
12. scores = model_selection.cross_val_score(clf, x, y,
    cv=0, scoring='accuracy')
13. Stop
```

a) XGBoost

Researchers have devised an algorithm called XGBoost for machine learning categorization that is incredibly effective. It is extremely quick, and its performance is improved because it is a boosted decision tree. This categorization prototype is used to boost the model's efficiency and speed.[28]

b) Random Forest Classifier

Leo Breiman of the University of California was the first to introduce random forest (RF) in 2001. It is composed of several simple classifiers (decision trees) which are independent of each other. A sample will be included in the new classifier and the class label of this sample depending on the voting outcomes from every single classification [29].

The main steps for the RF classifier are as follows:

1) Set the proper "M" value, which is the number of components of each sub-set of features.
2) Choose a new subset of feature HK from the entire feature set randomly based on the value of M. HK is free from another subset in h1; ...; HK sequence.
3) Training the data set for each training category with the feature sub-set to construct a decision tree. Every single category can be represented as h(X, hk) (where X specifies the inputs).
4) Select a new hk and repeat it until all the feature subsets are moving. An RF classifier has been achieved.
5) Input the test set. Decide based on voting outcomes for each classification of this sample.

3.7 Proposed Algorithm

Input: PID dataset

Output: Accuracy of categorization.

Approach:

Step 1. Initiate

Step 2. Obtain the UCI ML repository PID dataset.

Step 3. Preprocessed the collected dataset.

Step 4. Both zero values have been replaced with the median value of this attribute to clean the data set.

Step 5. In RFE, significant attributes are derived from the whole dataset to reduce data

Step 6. Scale the information using a standard scale.

Step 7. Data smoothing by a binning method including classification all features in bins UCI ML repository

Step 8. The Apriori method is used to identify common objects and lastly to generate association rules.

Step 9. Randomized Search CV with XGBoost5 and Random Forest Classifier for the adjustment of vote categorization parameters.

Step 10. a confusion matrix and different results

Step 11. Finish

Figure 1 describes a flow diagram of the proposed research project, with all essential phases listed from beginning to conclusion to provide a schematic view of the research.

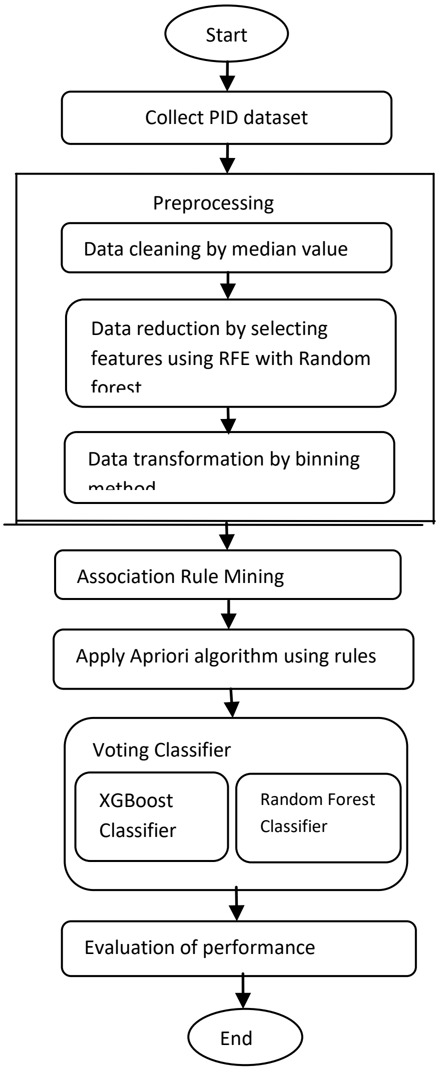
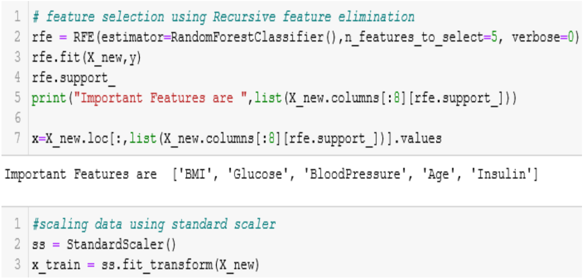

Start
Collect PID dataset
Preprocessing
Data cleaning by median value
Data reduction by selecting features using RFE with Random forest
Data transformation by binning method
Association Rule Mining
Apply Apriori algorithm using rules
Voting Classifier
XGBoost Classifier
Random Forest Classifier
Evaluation of performance
End

Fig -1: Flowchart of the Proposed Work

4. RESULT ANALYSIS

The proposed approach was tested using Jupyter Notebook Python programming as part of the research. The accuracy of the results and the ROC curve were used to assess them. A confusion matrix, as illustrated in figure 2, was used to estimate the accuracy of models. Experiments were conducted to fine-tune the model in terms of the number of decision trees and their depth.



```
# feature selection using Recursive feature elimination
rfe = RFE(estimator=RandomForestClassifier(),n_features_to_select=5, verbose=0)
rfe.fit(X_new,y)
rfe.support_
print("Important Features are ",list(X_new.columns[:8][rfe.support_]))

x=X_new.loc[:,list(X_new.columns[:8][rfe.support_])].values
```

```
Important Features are  ['BMI', 'Glucose', 'BloodPressure', 'Age', 'Insulin']
```

```
#scaling data using standard scaler
ss = StandardScaler()
x_train = ss.fit_transform(X_new)
```

Fig -1: Important feature selection Using RFE

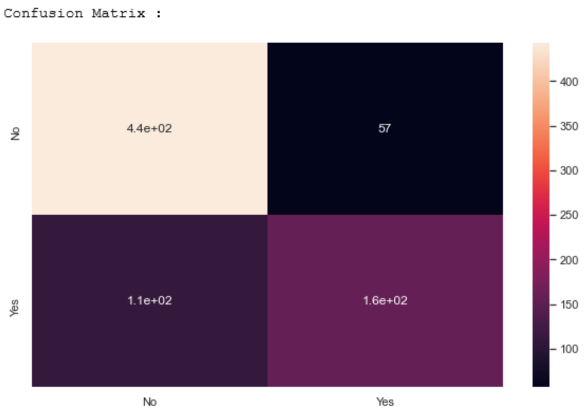


Fig -1: Confusion matrix of Random forest classifier

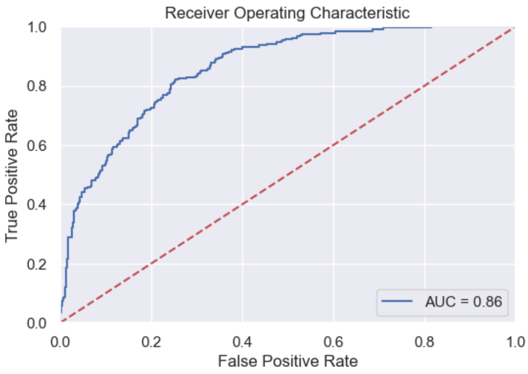
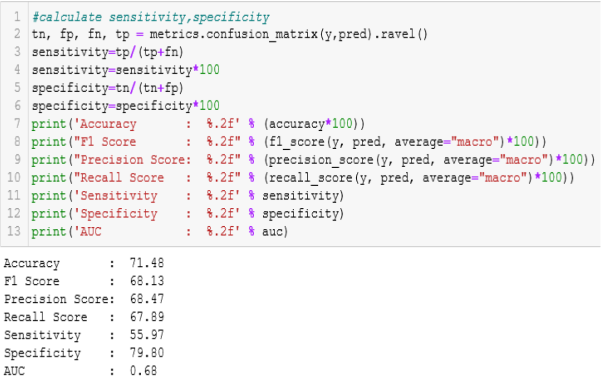


Fig -1: ROC curve of voting classifier

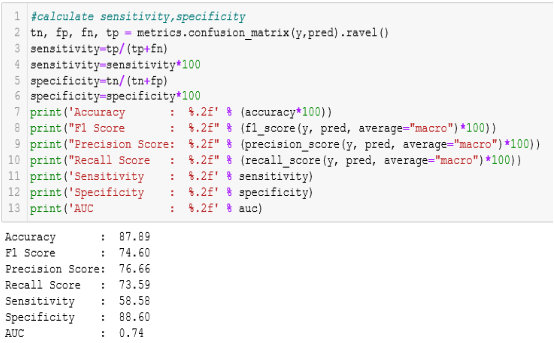


```
#calculate sensitivity,specificity
tn, fp, fn, tp = metrics.confusion_matrix(y,pred).ravel()
sensitivity=tp/(tp+fn)
sensitivity=sensitivity*100
specificity=tn/(tn+fp)
specificity=specificity*100
print('Accuracy       :  %.2f' % (accuracy*100))
print("F1 Score       :  %.2f" % (f1_score(y, pred, average="macro")*100))
print("Precision Score:  %.2f" % (precision_score(y, pred, average="macro")*100))
print("Recall Score   :  %.2f" % (recall_score(y, pred, average="macro")*100))
print('Sensitivity    :  %.2f' % sensitivity)
print('Specificity    :  %.2f' % specificity)
print('AUC            :  %.2f' % auc)
```

```
Accuracy       :  71.48
F1 Score       :  68.13
Precision Score:  68.47
Recall Score   :  67.89
Sensitivity    :  55.97
Specificity    :  79.80
AUC            :  0.68
```

Fig -1: Accuracy parameters for ANN classifier

Figures 5 and 6 are the result visualization of the values obtained by both ANN and voting classifier respectively which were run on Python.



```
#calculate sensitivity,specificity
tn, fp, fn, tp = metrics.confusion_matrix(y,pred).ravel()
sensitivity=tp/(tp+fn)
sensitivity=sensitivity*100
specificity=tn/(tn+fp)
specificity=specificity*100
print('Accuracy       :  %.2f' % (accuracy*100))
print("F1 Score       :  %.2f" % (f1_score(y, pred, average="macro")*100))
print("Precision Score:  %.2f" % (precision_score(y, pred, average="macro")*100))
print("Recall Score   :  %.2f" % (recall_score(y, pred, average="macro")*100))
print('Sensitivity    :  %.2f' % sensitivity)
print('Specificity    :  %.2f' % specificity)
print('AUC            :  %.2f' % auc)
```

```
Accuracy       :  87.89
F1 Score       :  74.60
Precision Score:  76.66
Recall Score   :  73.59
Sensitivity    :  58.58
Specificity    :  88.60
AUC            :  0.74
```

Fig -1: Accuracy parameters of the voting classifier

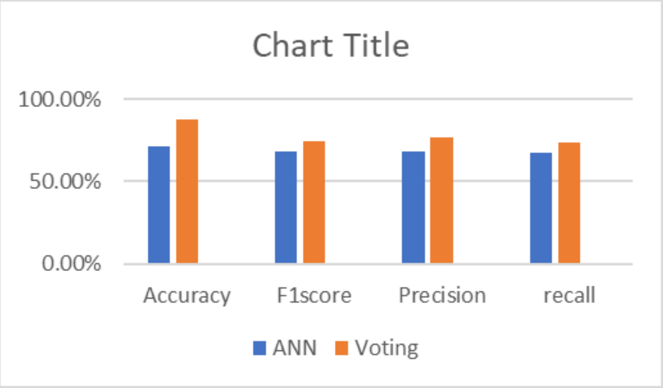


Fig -1: Graphical comparison of base and proposed methodology

The above figure shows the comparison of existing and current techniques using different parameters for evaluating the performance. The graph shows the higher values of all the considered parameters for a voting classifier.

5. CONCLUSIONS

In disease detection, machine learning and data mining techniques are critical. The ability to detect diabetes early on is critical in determining a patient's treatment options. The accuracy of a few existing diabetic medical diagnosis categorization systems is discussed in this research. In the expressions of accuracy, a classification issue has been discovered. The Pima Indians diabetes dataset was used to train and validate three machine learning techniques, which were then applied to a test dataset. In the future, the work can be expanded on a large dataset with more features and higher accuracy. We can also work on any clustering approach rather than a classification algorithm for visualizing the variations in the result.

REFERENCES

1. P. Radha and Dr. B. Srinivasan, “Predicting Diabetes by Sequencing the Various Data Mining Classification Techniques,” IJISET - International Journal of Innovative Science, Engineering & Technology Vol. 1 Issue 6, August 2014, pp. 334-339.
2. K.Priyadarshini and Dr. I.Lakshmi “A Survey on Prediction of Diabetes Using Data Mining Technique,” International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007

Certified Organization) Vol. 6, Special Issue 11, September 2017.

3. D Falvo and BE Holland, Medical and psychosocial aspects of chronic illness and disability. Jones & Bartlett Learning; 2017.
4. S. Skyler Jay, L. Bakris George, Ezio Bonifacio, Tamara Darsow, H. Eckel Robert, Leif Groop, Per-Henrik Groop, Yehuda Handelsman, A. Insel Richard, Chantal Mathieu, T. McElvaine Allison, P. Palmer Jerry, Alberto Pugliese, A. Schatz Desmond, M. Sosenko Jay, P.H. Wilding John and E. Ratner Robert, “Differentiation of diabetes by pathophysiology, natural history, and prognosis,” Diabetes, Vol. 66 Issue 2, Feb. 2017, pp. 241–255, doi.org/10.2337/db16-0806.
5. Z. Tao, A. Shi and J. Zhao, “Epidemiological Perspectives of Diabetes,” Cell Biochem Biophys, Vol. 73, February 2015, pp. 181-185, doi.org/10.1007/s12013-015-0598-4.
6. Sathya Chandrasekaran and K. Dharmarajan, “Survey on Data Mining Classification Techniques to Predict Diabetes”, Elysium journal, Vol. 4, Issue 4, August 2017, pp. 1-6.
7. Karnika Dwivedi and Dr. Hari Om Sharan, “Review on Prediction of Diabetes Mellitus using Data Mining Technique,” International Journal of Engineering and Technical Research (IJETR), ISSN: 2321-0869 (O) Vol. 8, Issue 12, December 2018 pp. 2454-4698.
8. R. Syed, R. K. Gupta, and N. Pathik, “An Advance Tree Adaptive Data Classification for the Diabetes Disease Prediction,” 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), July 2018, pp. 1793-1798, doi: 10.1109/ICRIEECE44171.2018.9009180.
9. Z. Xu and Z. Wang, “A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier,” 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), 2019, pp. 278-283, doi: 10.1109/ICACI.2019.8778622.
10. R. S. Raj, D. S. Sanjay, M. Kusuma, and S. Sampath, “Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes,” 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, pp. 41-45, doi: 10.1109/ICATIECE45860.2019.9063792.
11. S. A. Aboalnaser and H. R. Almohammadi, “Comprehensive Study of Diabetes Miletus Prediction Using Different Classification Algorithms,” 2019 12th International Conference on Developments in eSystems Engineering (DeSE), 2019, pp. 128-133, doi: 10.1109/DeSE.2019.00033.
12. R. Karthikeyan, P. Geetha and E. Ramaraj, "Rule-Based System for Better Prediction of Diabetes," 2019 3rd International Conference on Computing and Communications Technologies (ICCCT), 2019, pp. 195-203, doi: 10.1109/ICCCT2.2019.8824842.
13. Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, Muhammad Awais Malik, Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas, “A model for early prediction of diabetes,” Informatics in Medicine Unlocked, Vol. 16, 2019, 100204, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2019.100204.
14. M. Rout and A. Kaur, "Prediction of Diabetes Risk based on Machine Learning Techniques," 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020, pp. 246-251, doi: 10.1109/ICIEM48762.2020.9160276.
15. A. P, V. V. Nair, and N. S Nair, "Mellitus Preliminary Analysis using Various Data Mining Algorithms and Metrics," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1222-1225, doi: 10.1109/ICCES51350.2021.9489117.
16. S. Kumari and A. Singh, "A data mining approach for the diagnosis of diabetes mellitus," 2013 7th International Conference on Intelligent Systems and Control (ISCO), 2013, pp. 373-375, doi: 10.1109/ISCO.2013.6481182.
17. Fayssal Beloufa, and M A Chikh, "Design of Fuzzy Classifier for Diabetes Disease using Modified Artificial Bee Colony algorithm," Computer methods and programs in biomedicine, Vol. 112, Issue 1, Oct 2013, pp. 92-103, DOI: 10.1016/j.cmpb.2013.07.009
18. A. Methaila, P. Kansal, H. Arya and P. Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," Computer Science & Information Technology (CS & IT), 2014, pp. 53–59, DOI : 10.5121/csit.2014.4807.
19. B. Malley, D. Ramazzotti and J.T. Wu, “Data pre-processing,” Secondary analysis of electronic health records, Springer, Sep. 2016, pp. 115–41.
20. M. Egi, R. Bellomo, E. Stachowski, C. J. French, G. K. Hart, C. Hegarty and M. Bailey, "Blood glucose concentration and outcome of critical illness: the impact of diabetes," Critical care medicine, Vol. 36, Issue 8, Aug. 2008, pp. 2249-55, DOI: 10.1097/CCM.0b013e318181039a
21. M. Brunström and B. Carlberg, “Effect of antihypertensive treatment at different blood pressure levels in patients with diabetes mellitus: systematic review and meta-analyses,” weekly peer-reviewed medical trade journal (BMJ), Vol. 352, Issue 717, February 2016, DOI: https://doi.org/10.1136/bmj.i717
22. A. Menke, K.F. Rust, J. Fradkin, Y.J. Cheng and C.C. Cowie, “Associations between trends in race/ethnicity, aging, and body mass index with diabetes prevalence in the United States: a series of cross-sectional studies,” Annals of Internal Medicine, Vol. 161, Issue 5, Sep. 2014, pp. 328-35, DOI: 10.7326/M14-0286.
23. Z. Chiba, N. Abghour, K. Moussaid, A. El Omri and M. Rida, “A Novel Architecture Combined with Optimal

Parameters for Back Propagation Neural Networks Applied to Anomaly Network Intrusion Detection," Computers & Security, Vol. 75, June 2018, pp. 36-58.

24. N. Bindra and M. Sood, "Evaluating The Impact Of Feature Selection Methods On The Performance Of The Machine Learning Models In Detecting Ddos Attacks," Romanian Journal Of Information Science And Technology, Vol. 23, Issue 3, Jan. 2020, pp. 250–261.
25. J. Brownlee, "Recursive Feature Elimination (RFE) for Feature Selection in Python", Data Preparation, 2020.
26. S. K. Solanki and J. T. Patel, "A Survey on Association Rule Mining," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Feb. 2015, pp. 212-16, DOI: 10.1109/ACCT.2015.69.
27. https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/
28. K. Patil, S. D. Sawarkar and S. Narwane, "Designing a Model to Detect Diabetes using Machine Learning," International Journal of Engineering Research & Technology (IJERT), Vol. 08, Issue 11, Nov. 2019, pp. 333-340.
29. A. Parmar, R. Katariya and V. Patel, "A Review on Random Forest: An Ensemble Classifier," Lecture Notes on Data Engineering and Communications Technologies, Vol. 26, Dec. 2018, pp. 758–763, doi:10.1007/978-3-030-03146-6_86.

BIOGRAPHIES

Rajasekhar c received the B. Tech degree in Computer Science and Engineering from Mahatma Gandhi University, India in 2011. He is Currently pursuing M. Tech degree in Computer Science and Engineering from Mount Zion College of Engineering, Kadammanitta, Kerala, India. His primary research interests are in Networking and Artificial Intelligence (Machine Learning Oriented Programming).

Bibin Varghese is Currently working as Assistant Professor in the Department of Computer Science and Engineering at Mount Zion College of Engineering, Kadammanitta, Kerala, India. His primary research interests are in Cyber Security and Artificial Intelligence.

Dr Smita C Thomas is Currently working as Associate Professor in the Department of Computer Science and Engineering at Mount Zion College of Engineering, Kadammanitta, Kerala, India. Her primary research interests are in cloud computing, Image Processing, Data Mining, Cyber Security and Artificial Intelligence (Machine Learning Oriented Programming).