iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46567

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com

Diabetes Prediction Using Different Ensemble Learning Classifiers in Machine Learning

P. Manoj Kumar[1], K.V. S Haswanth[2], G. Mahidhar Swaroop[3], M. Jasmine Pemeena Priyadarsini[4]
[1, 2, 3, 4]*School of Electronics Engineering, Vellore Institute of Technology, Vellore, 632014, Tamil Nadu, India*

Abstract: ***Considered a chronic illness, Diabetes results due to increased level of the glucose in blood in the body which happens due to either less insulin production or if the response to insulin by body cells is not proper. Current practice in hospitals is to collect the required information for diabetes diagnosis through various tests and treatment is given based on the test results. Producing accurate results through prediction models of diabetes is difficult because there is not much data available and there is presence of outliers as well. This Literature proposes an optimal prediction model for diabetes where the raw data collected will go through few pre-processing techniques before introducing to the ML Classifiers such as Random Forest, AdaBoost, XGBoost. Using the pre-processing techniques and ensemble methods we have got better performance results. The weights of ML models are reviewed using their respective Area Under ROC Curve (AUC) result.***
Keywords: Machine Learning, Classification, Pima Indian Diabetic dataset, missing values, outlier rejection, Random forest, AdaBoost, XGBoost classifier.

I. INTRODUCTION

Diabetes mellitus can be broadly classified as Type 1 diabetes, Type 2 diabetes and gestational diabetes. Where, occurrence of Type 1 diabetes happens when the cells in our pancreas that produce insulin are mistakenly attacked by our immune system. And is common among children. Type 2 is most common diabetic condition in India. Here, the body can't use insulin efficiently. Therefore, the pancreas start producing insulin in more amounts leading to high blood sugar. And gestational diabetes occurs in pregnant women due to insulin blocking hormones produced during pregnancy. Therefore, occurs only during pregnancy. There is no cure for Type 1 diabetes but type 2 diabetic condition is easy to avoid by doing regular exercise, controlling diet to reduce weight, not smoking and maintaining low cholesterol levels. So, prediction of diabetes in prior has been in the spotlight of researchers worldwide. Technologies that help scientist in predicting diabetes include Big Data Analytics, ML and Data mining. These are all interlinked and are trending approach that are used to solve real time problems.

For the prediction analysis we are using PIMA Indians diabetes data set provided by National Institute of diabetes and digestive and kidney diseases. Firstly, the obtained data is pre-processed using certain methods mentioned in (III C) before providing it to the ML models. In comparison with the previous researches (II) on diabetes prediction, we have improved the result by Training and testing using Shuffle split cross validation method and each algorithm is tuned by using a greater number of parameters in the grid search CV method. To implement the machine learning algorithms, we are using Google Colab as ide. Considerable number of experiments on various combinations of pre-processing methods and feature selection on ML models are carried out in order to find the most accurate ensemble classifier, which uses the best results yielding pre-processing methods and hyperparameters from experiments. At the end of this paper will be able to predict the diabetes from the data given with the highest accuracy possible. Here, we are using The Area Under the Curve AUC as evaluating parameter as it best describes the ROC curve that measures the ability of the classifier to differentiate between classes and is unbiased to the class distribution.

II. LITERATURE SURVEY

The authors in [1] have put in efforts to implement both the support vector machine (SVM) and Naïve Bayes statistical model combinedly for predicting diabetes. [3] informs about how prediction models vary with inclusion of feature selection method like principal component analysis. Ensemble boosting algorithm has been used in [5] in predicting undiagnosed people. The authors in [8] have discussed about the feature importance in predicting the diabetes through Support Vector Machines (SVM), Random Forest and Logistic Regression classifier. Researchers in [9] have proposed taking partitioning based on tree as an advantage, and classification based on adaptive SVM approach. The authors in [10] have discussed about different types of diabetes and causes for it. And they used different ML model like Naive Bayes (NB), Decision Tree, SVM algorithms for classification. [11] has detailed implementation of random forest classification algorithm for diabetes prediction.

The researchers in [12] have analysed preliminary prediction of diabetes by utilizing various factors related to this disease using ML techniques, namely K- Nearest Neighbour (KNN), C4.5 decision tree, Support Vector Machine (SVM), Naive Bayes (NB). The authors in [13] have proposed a framework that uses ensemble models and neural networks for predicting diabetes. AdaBoost algorithm has been used in [14] with various ML algorithms as a base classifier for predicting Diabetes Mellitus-A. The authors in [19] have proposed a ensemble model by involving various ML models like Logistic Regression Classifier, Decision Tree, KNN, AdaBoost, Random Forest for diabetes retinopathy classification.

III.MATERIALS AND METHODS

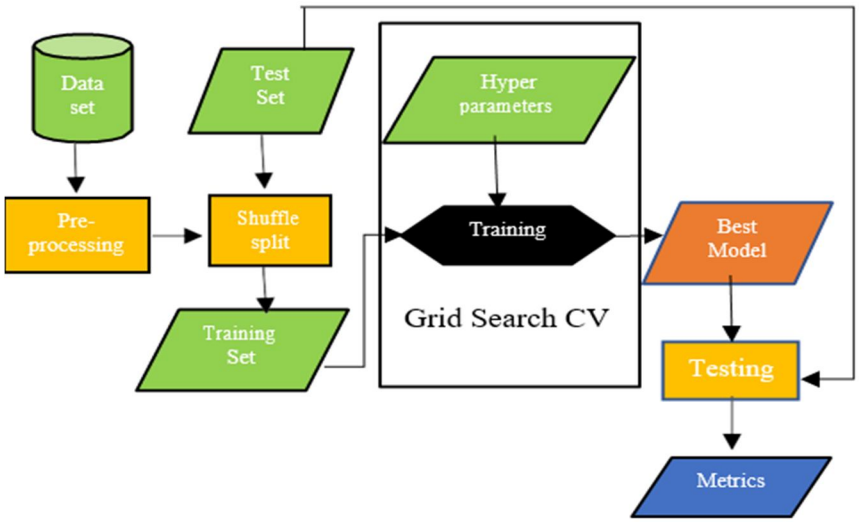


Fig. 1 Architecture of proposed prediction system

A. Data set

The prediction analysis is made using the data set created by the National Institute of Diabetes and Digestive and Kidney Diseases. Motive behind creating this data set is to predict Diagnostically if the patient is diabetic or not. And the deductions will be made considering certain Diagnostic measurements provided in the data set. Dataset is created aiming to determine if a patient is diabetic or not using the diagnostic measurements present in the dataset. And data set consists of all females aged 21 or above. Few predictor variables included are pregnancies count, BMI, glucose level, age and few more. And there is One outcome variable that determines whether patient is diabetic or not and is represented by 1 and 0 respectively.

B. Diabetes Pedigree Function

"A synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject," is what represented by diabetes pedigree feature. It takes inputs from the data of bloodline of a person to estimate how they are affected by diabetes.

$$\frac{\sum_m I_m(88-ADM_m)+20}{\sum_n I_n(ALC_n-14)+50} \quad (1)$$

here m and n respectively denote the relatives who are diabetic and not diabetic. I denotes the percentage of genes shared by the relatives (I = 0.500 corresponds to parent and full sibling, I = 0.250 corresponds to half-sibling, grandparent, aunt, uncle and I= 0.125 for a half aunt or half-uncle and first cousin). $ADM_m$and ACL_n is the relatives age in years during diagnosing period and when the last test that resulted non-diabetic respectively.

C. Pre-Processing Techniques

Data pre-processing is an important step in Machine Learning because the quality of data and the valuable knowledge that can be obtained from it directly impacts the ability of model to learn; thus, pre- processing our data before feeding it into our model is critical.

TABLE I
DESCRIPTION OF ATTRIBUTES MENTIONED IN DATA SET

S.NO	Attribute	Description	Values Range	Mean
1	Pregnancies	Number of pregnancies	0-17	3.845052
2	Glucose	Glucose Concentration	0-199	120.894531
3	Blood Pressure	Blood pressure	0-122	69.105469
4	Skin Thickness	skin fold thickness of triceps	0-99	20.536458
5	Insulin	2-hour serum insulin	0-846	79.799479
6	BMI	Body mass index	0-67	31.992578
7	DPF	Diabetes pedigree function	0-2.45	0.471876
8	Age	Age of an individual	21-81	33.240885
9	Class	Tested positive/negative	0 and 1	0.348958

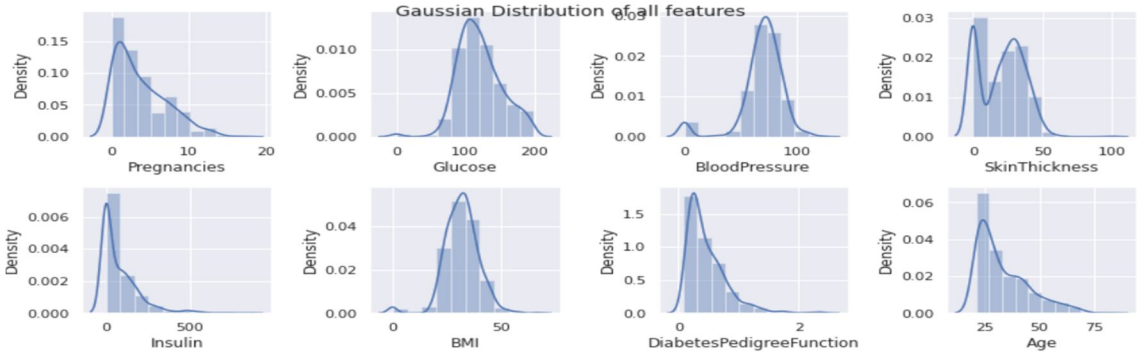


Fig. 2 Gaussian distribution of features in the dataset.

1) Outlier Rejection

Outliers are the abnormal observations that are different from other observations in the attributes. As ML model are data sensitive, so there is need for rejecting these observations. Based on Interquartile Range score outliers are detected and rejected in our work. Steps involved in IQR calculation is as follows:

- Calculating third quartile-Q3(75 percentile value), first quartile-Q1(25 percentile value).
- Calculating IQR=Q3-Q1.
- Finding Lower Bound & Upper Bound

(LB)= Q1 – 1.5 * IQR
(UB)= Q3 +1.5 * IQR.

If observation lies between Lower Bound (LB) and Upper Bound (UB) then it is not treated as outlier and value will not be rejected. Otherwise, value will be rejected.

2) Filling missing values

Many datasets may contain missing values in them. To fill those missing values instead of rejecting them we are taking mean of that attribute and then we replace missing values with the mean value.

$$\text{Mean} = \frac{sum\ of\ all\ values\ in\ the\ attribute}{No.of\ values\ in\ the\ attribute} \quad (2)$$

3) Standardization

Standardization is the technique to rescale all the attributes values into common scale. And also, for achieving standard normal distribution with unit variance and zero mean. The Standardization as follows.

$$S(x) = \frac{x - mean(x)}{standard\ deviation\ of\ x} \quad (3)$$

4) Feature Selection

Feature Selection is the technique to select the most relevant features that contribute better output. To achieve this, we are using correlation-based feature selection. Correlation is the measure of how much two random variables X and Y are linearly correlated. the correlation formula is as follows:

$$F = \frac{\sum(X_i - mean(X))(Y_i - mean(Y))}{\sigma_X * \sigma_Y} \quad (4)$$

Sort the F values in ascending order to select first k features.

IV. MODEL TRAINING AND VALIDATION

A training model is a dataset used to train a machine learning algorithm. It is made up of sample output data as well as the related sets of data given as input that effects the output. Training models are to process the input data through the algorithm and then compare the resulted output to the output sample in dataset. The model is modified based on the results of this correlation.

A. Shuffle split cross-validation

The Shuffle Split iterator can create as many separate train / tests dataset splits as the user specifies. The samples are shuffled before being divided into two train and test sets. The purpose and difference of using shuffle split instead of k-fold cross validation is that K- Fold divides the data set into a predetermined number of folds, with each sample belonging to just one-fold. During each iteration, Shuffle Split will randomly sample your entire dataset to create a training set and a test set. Since you are sampling from the entire dataset during each iteration, values chosen in one iteration will be chosen again in a subsequent iteration.

In K-Fold, one-fold is used as the test set and the remaining folds as the training set during each round. However, in Shuffle Split, you can only use the training and test sets from iteration n during each round n. Cross validation time increases as the data set expands, making shuffle splits a more appealing alternative. If you can train your algorithm with a percentage of your data, that's great. Shuffle Split is an appealing choice if you can train your algorithm with a portion of your data rather than using all k-1 folds.

B. Hyperparameter Tuning

The process of using different combinations of hyperparameters for a learning algorithm in order to find the set that gives best results is known as hyperparameter tuning in machine learning. A hyperparameter is a value for a parameter that is used to guide the learning process. Other parameters, such as node weights, are, on the other hand, learned.

To generalise different data patterns, the same prediction model may require different parameters, weights, or learning speeds. These parameters are known as hyperparameters, and they must be fine-tuned in order for the model to solve the prediction problem optimally. Hyperparameter optimization identifies a set of hyperparameters that results in the best model that minimizes any loss. GridSearchCV and RandomizedSearchCV are two basic methods for Hyperparameter optimization. Grid search is the most common form of hyperparameter tuning. Here, using this technique we create a model for each possible combination of all hyperparameter values given, validate each model, and select the combination that gives us the best results. As, the number of parameters is less and we don't want to miss out on any combination, we prefer Grid search as our Hyperparameter optimization method.

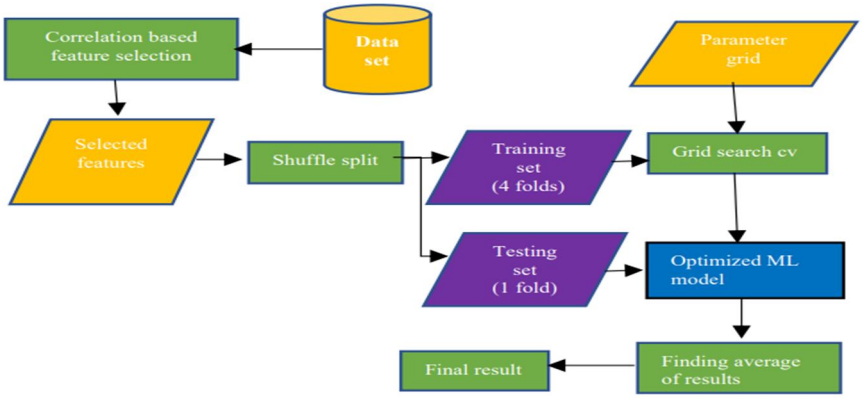


Fig. 3 cross validation with hyperparameter tuning using GridSearchCV

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com

TABLE II
LIST OF HYPERPARAMETERS USED TO TUNE EACH ML MODEL USING GRIDSEARCHCV

ML Model	Hyperparameters used
Random Forest (RF)	1. Function to measure quality of split (Gini impurity, entropy) 2. Minimum samples to split internal node 3. Minimum samples at leaf node
AdaBoost (AB)	1. The algorithm for boosting (either real or discrete) ('SAMME', 'SAMME.R') 2. Learning rate to shrink the contribution of each classifier 3. The maximum number of estimators to terminate the boosting
XGBoost (XG)	1. instance weight (Hessian) Minimum sum in child 2. For further partitioning on the leaf node required minimum loss reduction 3. Ratio of subsample for each training instance 4. Ratio of subsample of columns while construction of each tree 5. Maximum possible depth of a tree

V. ALGORITHMS

A. Random Forest Algorithm

In simpler terms it is process of solving a complex problem by combining multiple classifiers which results in increase in performance of the model or it can be defined as a classifier containing a number of decision trees in multiple subsets of the given data which takes the average of decision trees to improve accuracy of the model. Random Forest almost has same hyperparameters as a bagging classifier or a decision tree. Each tree in the Random Forest predicts a class and predicted class gaining major votes is declared as our model prediction.

The main importance of the Random Forest algorithm is to recognize the most important features of a given dataset. However, there are a few disadvantages apart from the advantages i.e., the complexity is one of the main disadvantages and building a random forest-based prediction model is difficult and time consuming than a decision tree-based model, it requires more resources in order to work effectively.

B. ADABOOST

AdaBoost is a statistical classification meta-algorithm. It can be combined with various learning algorithms to boost results. All weak learners' output is combined to give a weighted sum that represents the boosted classifier's output. AdaBoost is adaptive in nature that it tweaks further poor learners with help of weighted sum in favour of output misclassified by previous weak learners.

In certain situations, it is rarely subjected to overfitting problem than other learning algorithms. Weak learners will be poor, considering until their output is slightly better than random guessing, the final model will transform to a strong learner.

Every learning algorithm has several parameters to modify before it achieves best resulting output for a dataset, and most of them fits certain problem types better than others. Often AdaBoost is considered to be the strongest out-of-the-box classifier (with the weak learners comprising of decision trees).

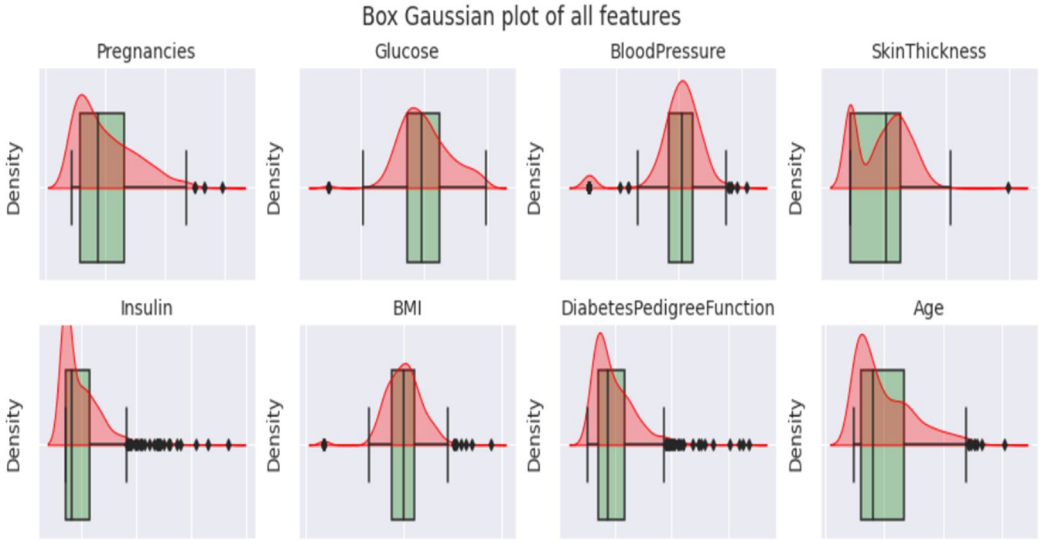
C. XGBoost

XGBoost is additionally referred to as Extreme gradient boosting technique. Ensembling is a type of learning process in which training of multiple ML models takes place and result in optimised predictions improving the performance of single ML model. It is one of the boosting techniques in ensemble learning which aims to build a strong classifier from weak learners. Extreme Gradient boosting is a method where the new models are created that finds the error in the previous model and then residue is added to make the final prediction. XGBoost carries out gradient boosting decision tree algorithm. XGBoost provides a fast implementation of the stochastic gradient boosting algorithm as well as access to a set of model hyperparameters for fine-tuning the model training process.

VI. RESULTS AND DISCUSSION

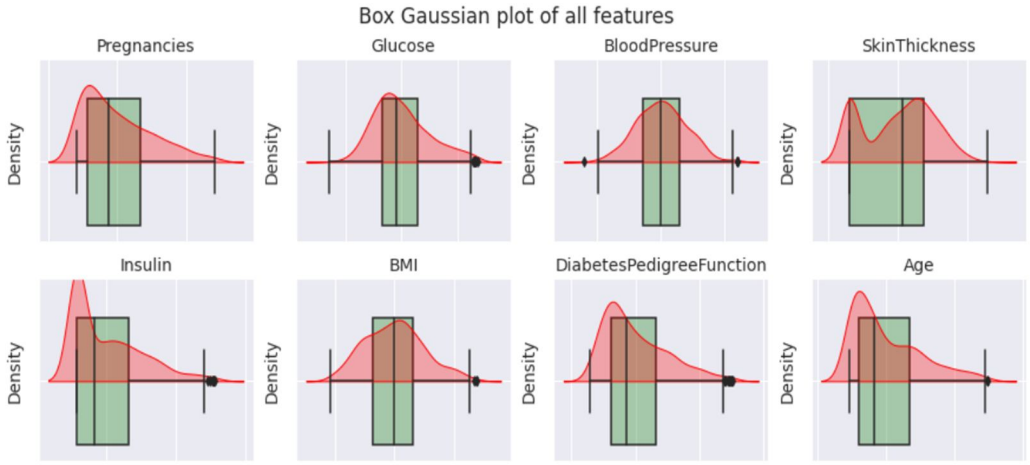
A. Pre-Processing Results

Most of the data of the attributes in the Dataset is incomplete due to presence of the outliers and missing values so the data has to be processed through certain pre-processing techniques so that the ML model can outrun the incompetence of the dataset and can be trained for better results. However, because of the presence of outliers in the data it introduces kurtosis and skewness in the distribution of attribute's **Fig.4(a)** and high kurtosis is an indication for presence of a greater number of outliers in the data. Skewness and Kurtosis will affect the result in such a way that will lead to underestimation and overestimation of the expected value respectively. As demonstrated in the **(Fig (4))** result of outlier rejection of the dataset the skewness of the distribution is moved to zero means **(Fig.4(b))** indicating attribute's mean value and median value have almost coincided. Confusion matrix mentioned below **(Fig.5)** shows the results of the raw data after going through the pre-processing techniques i.e., results of the data after removal of outliers and replacing missing values. Figures **Fig.5(a) and Fig.5(b**) shows the qualitative and quantitative analysis of the increase in correlation coefficient implying that the correlation between attribute and the target outcome has been improved after the data has been processed by removal of outliers and missing values replaced by the mean of the attribute they belong to. when we observe the results closely removal of outliers and replacing missing values helped to improve the correlation coefficient for the F3, F4 and F5 with respect to outcome. This improvement in correlation helps us to select the most correlated attributes in feature selection.



(a) When outliers are present in the data

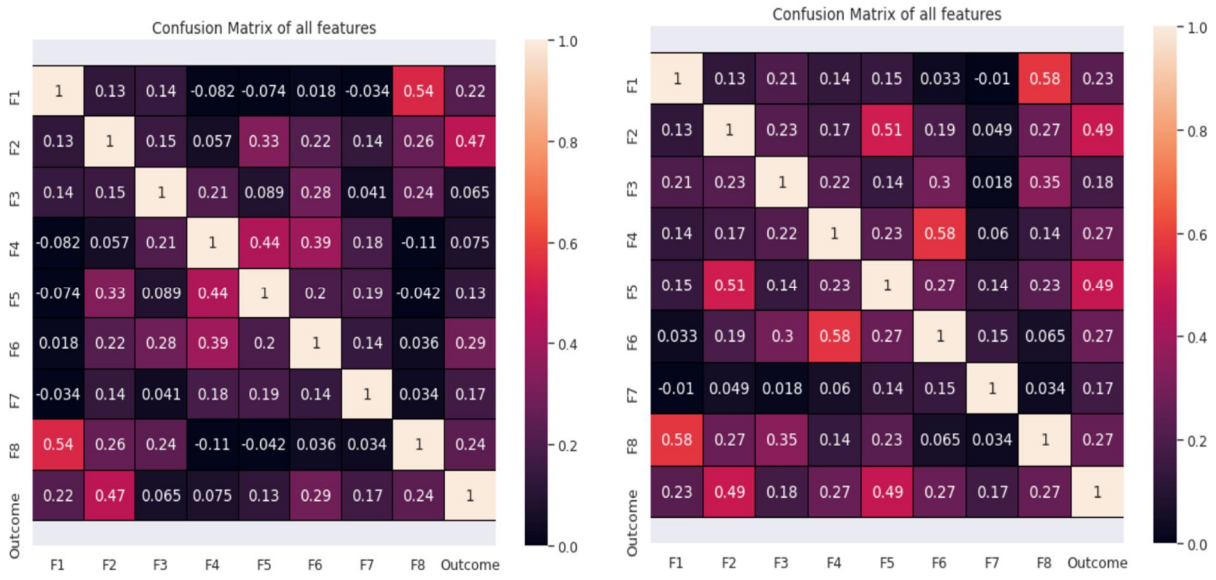
(b)



(c) After outliers are rejected in the data

Fig. 4 Each attribute's Gaussian boxplot (a)with outliers and (b) without outliers.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com



(a) Before pre-processing (b) After pre-processing

Fig. 5 Each attribute's correlation with the outcome as confusion matrix for (a) raw data and (b) pre-processed data

VII. OVERALL EXPERIMENTAL RESULTS

Table III displays the results for selecting the optimal performing pre-processing methods combination and ML algorithm, with AUC recorded to compare among them. Table IV summarizes each model's ability to achieve the highest AUC following the proposed framework, together with the best pre-processing methods and feature selection algorithm and the number of features selected. Table IV also includes the best-tuned hyperparameters obtained via grid quest. Table III shows that when we use enough pre-processing, we can get better results from various models.

Table III demonstrates the classification efficiency has improved substantially when missing values are replaced(B) with mean of the attribute it belongs to, rather than rejection and removal of outliers (A). When both A and B are used, the XB has prevailed in any case of function selection. The addition of standardization furtherly as a pre-processing step does not boost the classifiers' efficiency because it is not always guaranteed to do so. Therefore, standardization couldn't help to increase the efficiency of certain ML models in this literature (Table III).

Table IV also shows that most classifiers performed better with six attributes than with four. From feature selection we can conclude that the role of certain attributes such as Blood Pressure(F3) and diabetes pedigree function(F7) can be ignored. since they bear less diabetes detail in them of the PID dataset for diabetes prediction in comparison to other features.

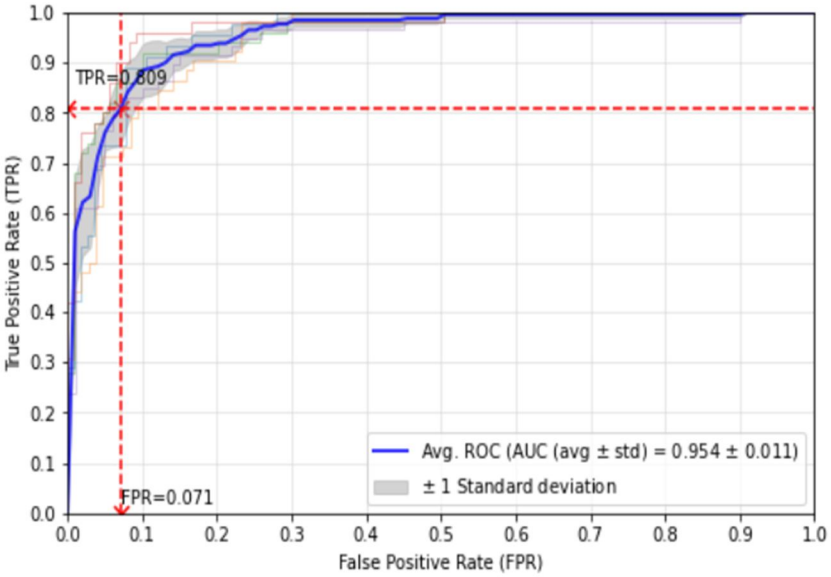
TABLE III

Experimental results of best performing pre-processing methods for each algorithm to get highest possible AUC. Where (A) is outlier rejection, (B) is filling missing values, (C) is standardisation

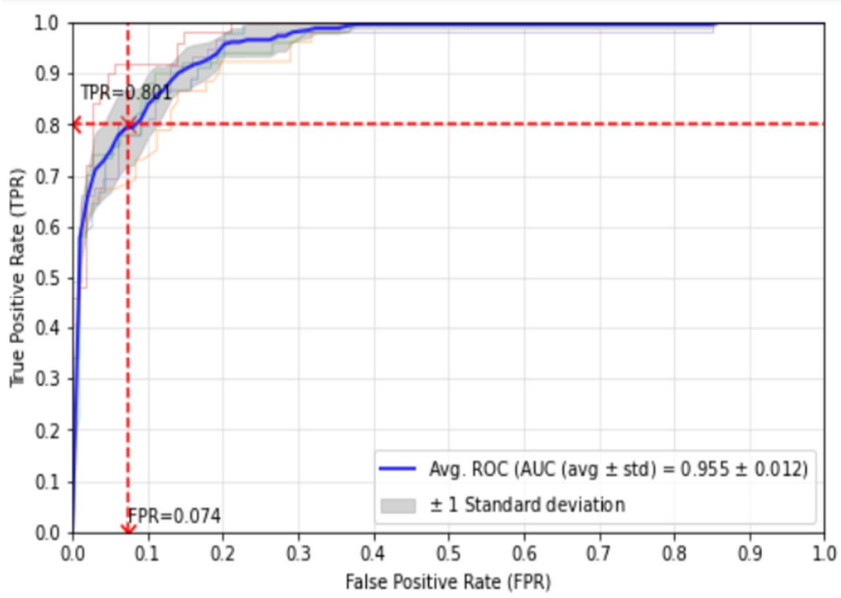
Pre-Processing	Feature Selection method	N	RF	AB	XG	Best
A	Correlation	4	0.798 +/- 0.020	0.796 +/- 0.021	0.805 +/- 0.018	XG
		6	0.832 +/- 0.012	0.831+/- 0.016	0.829 +/- 0.014	RF
A+B	Correlation	4	0.951+/- 0.012	0.952 +/- 0.010	0.954 +/- 0.007	XG
		6	0.953 +/- 0.011	0.955 +/- 0.012	0.961 +/- 0.014	XG
A+B+C	Correlation	4	0.951+/- 0.012	0.952 +/- 0.010	0.955 +/- 0.008	XG
		6	0.954 +/- 0.011	0.955 +/- 0.012	0.960 +/- 0.014	XG

TABLE IV
BEST RESULT GIVING HYPERPARAMETERS USED FOR ML MODELS THROUGH GRIDSEARCHCV

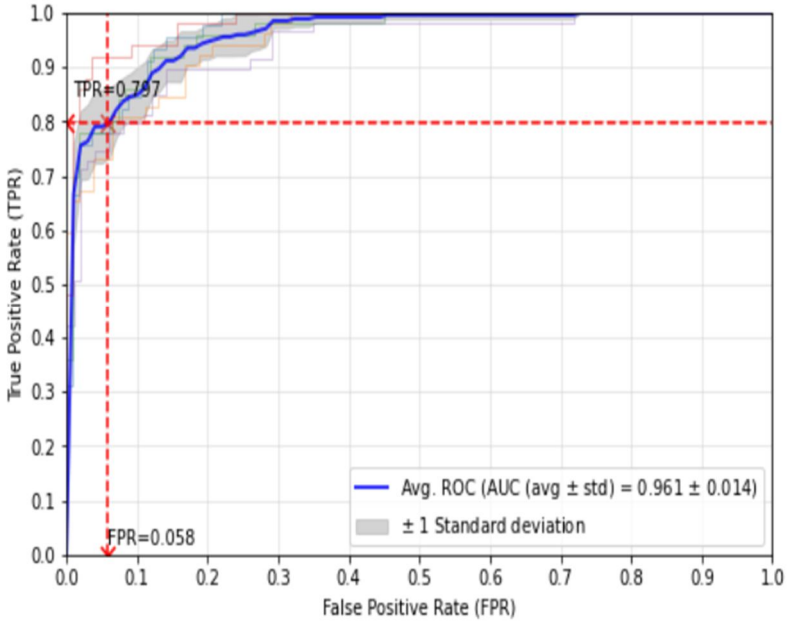
Classifier	Best pre-processing	Best Hyperparameters	AUC result
RF	A+B+C and Correlation (attributes=6)	{'min_samples_split': 0.2, 'min_samples_leaf': 4, 'criterion': 'entropy'}	0.954 +/- 0.011
AB	A+B+C and Correlation (attributes=6)	{'n_estimators': 200, 'learning_rate': 1.0, 'algorithm': 'SAMME'}	0.955 +/- 0.012
XG	A+B and Correlation (attributes=6)	{'subsample': 1.0, 'min_child_weight': 5, 'max_depth': 5, 'gamma': 1.5, 'colsample_bytree': 0.6}	0.961 +/- 0.014



(a) Random Forest



(b) AdaBoost



(c) XGBoost

(d)

	Condition positive	Condition negative	
Predicted positive	TP 204	FP 31	Precision 0.866
Predicted negative	FN 52	TN 508	FOR 0.093
	SN 0.798	SP 0.942	

(d) Confusion matrix of XGBoost

Fig. 6 AUC result of all the ML models and confusion matrix of XGBoost is given as example

VIII. EVALUATION METRICS

All of the detailed experiments were evaluated using a variety of metrics, each with its own definition of evaluation. The True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) of confusion matrix has been included (as shown in **Fig.6 (d))**, as well as various metrics such as Sensitivity (Sn) (5), Specificity (Sp) (6), Precision (Pr) (7), False Omission Rate (FOR) (8), and Diagnostic Odds Ratio (DOR) (9) are also included.

Sensitivity (Sn) = $\frac{TN}{TP+FN}$ (5)

The Sensitivity and Specificity are used to measure type-II error (when a patient has positive symptoms but is incorrectly rejected) and type-I error (when a patient has negative symptoms but is incorrectly detected as positive).

Specificity (Sp) = $\frac{TN}{FP+TN}$ (6)

Pr, FOR, and DOR have been used to assess the proportion of correctly diagnosed diabetes patients with positive conditions, the proportion of people who have a negative test result but have a positive true diagnosis, and the diagnostic test's efficacy, respectively.

$$\text{Precision (Pr)} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{False Omission Rate (FOR)} = \frac{FN}{FN+TN} \quad (8)$$

The diagnostic odds ratio (DOR) is an indicator of a diagnostic test's efficacy in medical research with binary classification. It is known as the ratio of the chances of a positive test if the subject has a disease to the chances of a positive test if the subject does not have the disease.

$$\text{Diagnostic Odds Ratio (DOR)} = \frac{TP*TN}{FP*FN} \quad (9)$$

Additionally, rather than reporting absolute values, the Receiver Operating Characteristics (ROC) with Area Under the ROC Curve (AUC) is used to assess how well predictions are ranked.

IX. CONCLUSION

The main objective of this study is to propose a promising ensemble ML model for the prediction of diabetes at an early stage, the data we have used is the PIMA dataset, due to the incomplete data the raw data has been introduced to effective pre-processing techniques which outlines the factors of presence of outliers, missing values and makes the data standardized before introducing to the ML models. By using shuffle split instead of the k-folds the accuracy have model has been increased and got better results, weights of the ML models have been measured and compared by the performance metric Area Under Curve(AUC).The ensemble models we have used are Random Forest, AdaBoost, XGBoost and their results are **0.954 +/- 0.011, 0.955 +/- 0.012, 0.961 +/- 0.014** (from figure.6) respectively after the training and testing by comparison of these results and model evaluation we propose the model XGBoost with accuracy **0.961 +/- 0.014** as the best model for the prediction of diabetes at an early stage. And comparison of other metrics is provided in Table V.

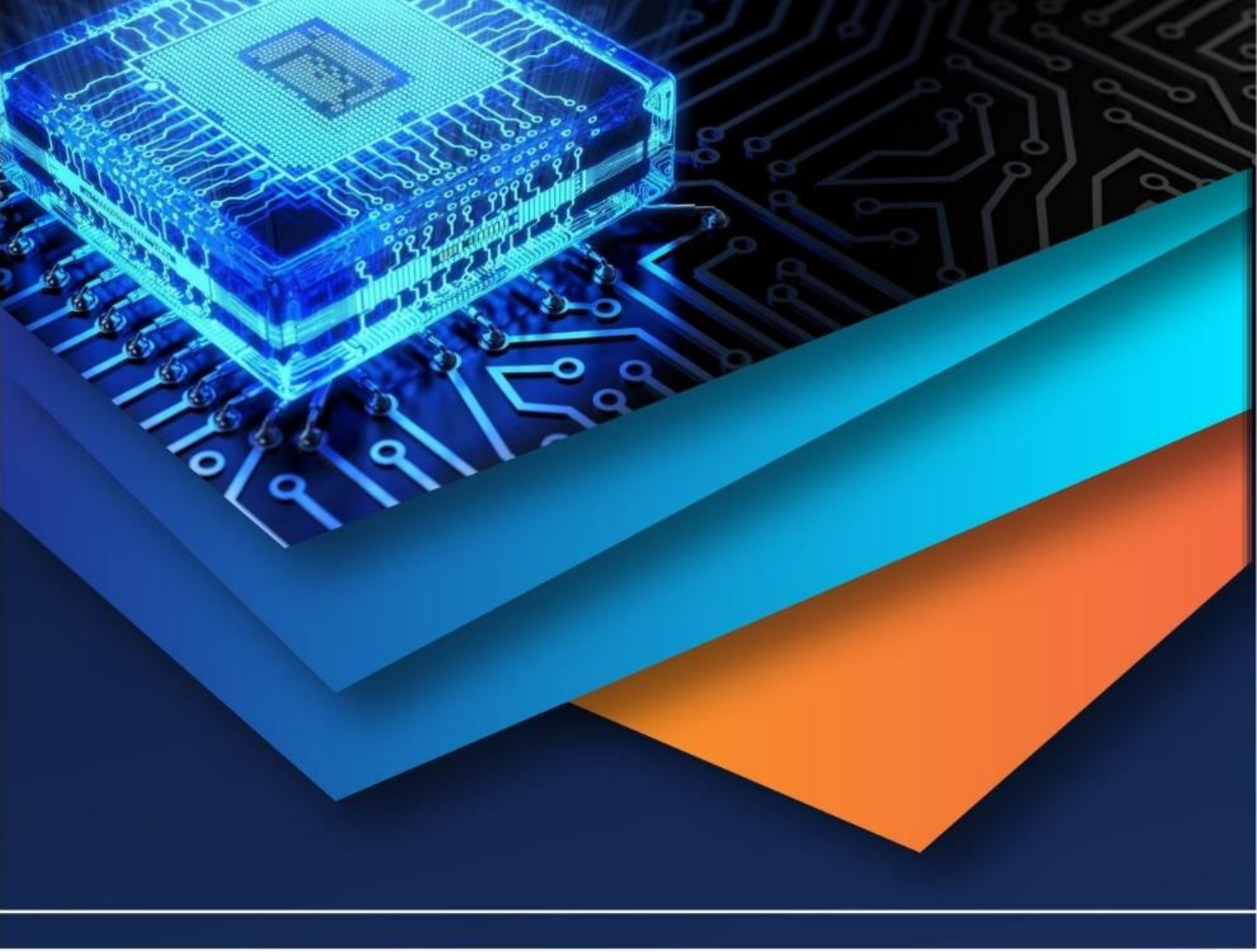
TABLE V
COMPARISON AMONG THE ALGORITHMS ON DIFFERENT EVALUATION METRICS MENTIONED

Metrics	RF	AB	XG
Precision	84.5%	84%	86.6%
Sensitivity	81%	80.2%	79.8%
Specificity	93%	92.6%	94.2%
Accuracy	89.1%	88.6%	89.6%
FOR	89%	93%	93%
DOR	64.162	69.53	89.55
AUC	95.4%	95.5%	96.1%

REFERENCES

[1] Zhilbert Tafa, Nerxhivane Pervetica, and Bertran Karahoda. An Intelligent System for Diabetes Prediction. 2015 4th Mediterranean Conference on Embedded Computing (MECO).
[2] Aparimita Swain, Sachi Nandan Mohanty, and Ananta Chandra Das.Comparative risk analysis on prediction of Diabetes Mellitus using machine learning approach. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).
[3] B. Dhomse Kanchan and M. Mahale Kishor. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis.2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.
[4] Ridam Pal, Dr.Jayanta Poray, and Mainak Sen. Application of Machine Learning Algorithms on Diabetic Retinopathy. 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).
[5] Roxana Mirshahvalad and Nastaran Asadi Zanjani. Diabetes Prediction Using Ensemble Perceptron Algorithm. 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN).
[6] Deeraj Shetty, Kishor Rit, Sohail Shaikh, and Nikita Patil.Diabetes Disease Prediction Using Data Mining. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).
[7] Ayman Mir and Sudhir N. Dhage. Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
[8] Debadri Dutta, Debpriyo Paul and Parthajeet Ghosh. Analysing Feature Importances for Diabetes Prediction using Machine Learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).
[9] Rukhsar Syed, Rajeev Kumar Gupta, and Nikhlesh Pathik. An Advance Tree Adaptive Data Classification for the Diabetes Disease Prediction. 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE).

[10] Priyanka Sonar and Prof. K. JayaMalini. Diabetes Prediction Using Different Machine Learning Approaches. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).
[11] K. VijiyaKumar, B. Lavanya, I. Nirmala,and S. Sofia Caroline. Random Forest Algorithm for the Prediction of Diabetes. 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN).
[12] Md. Faisal Faruque, Asaduzzaman and Iqbal H. Sarker. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE).
[13] Md. Kamrul Hasan,Md. Ashraful Alam, Dola Das, Eklas Hossain and Mahmudul Hasan. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access (Volume: 8)
[14] Veena Vijayan V and Anjali C.Prediction and Diagnosis of Diabetes Mellitus-AMachine Learning Approach .2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS).
[15] Muhammad Azeem Sarwar,Nasir Kamal,WajeehaHamid, Munam Ali Shah.Prediction of Diabetes Using Machine Learning Algorithms in Healthcare.2018 24th International Conference on Automation and Computing (ICAC).
[16] Gulam GausWarsi, SoniaSaini, and Kumar Khatri. Ensemble Learning on Diabetes Data Set and Early Diabetes Prediction.2019 International Conference on Computing, Power and Communication Technologies (GUCON).
[17] Shon Mathew jacob, Kumudharaimond and Deepak anmani.Associated Machine Learning Techniques based On Diabetes Based Predictions.2019 International Conference on Intelligent Computing and Control Systems (ICCS).
[18] Adil Laabidi and Mohammed Aissaoui. Performance analysis of Machine learning classifiers for predicting diabetes and prostate cancer.2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)
[19] Reddy, G.T, Bhattacharya, S., Siva Ramakrishnan, S., Chowdhary, C. L., Hakak, S., Kaluri, R., & Praveen Kumar Reddy, M.An Ensemble based Machine Learning model for Diabetic Retinopathy Classification.2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE).
[20] 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, doi: 10.1109/ACCESS.2019.2929866.
[21] E. A. Pustozerov et al., "Machine Learning Approach for Postprandial Blood Glucose Prediction in Gestational Diabetes Mellitus," in IEEE Access, vol. 8, pp. 219308-219321, 2020, doi: 10.1109/ACCESS.2020.3042483.
[22] N. Barakat, A. P. Bradley and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pp. 1114-1120, July 2010, doi: 10.1109/TITB.2009.2039485.
[23] E. Montaser, J. -L. Díez, P. Rossetti, M. Rashid, A. Cinar and J. Bondia, "Seasonal Local Models for Glucose Prediction in Type 1 Diabetes," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 2064-2072, July 2020, doi: 10.1109/JBHI.2019.2956704.
[24] K. Li, C. Liu, T. Zhu, P. Herrero and P. Georgiou, "GluNet: A Deep Learning Framework for Accurate Glucose Forecasting," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 414-423, Feb. 2020, doi: 10.1109/JBHI.2019.2931842.
[25] S. P. Chatrati, G. Hossain, A. Goyal, A. Bhan, S. Bhattacharya, D. Gaurav, and S. M. Tiwari, ‘‘Smart home health monitoring system for predicting type 2 diabetes and hypertension,’’ J. King Saud Univ. Comput. Inf. Sci., Jan. 2020.
[26] L. Jiang, Z. Cai, D. Wang, and S. Jiang, ‘‘Survey of improving K-Nearest Neighbor for classification,’’ in Proc. 4th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD), Aug. 2007, pp. 679–683.
[27] R. E. Schapire, ‘‘Explaining AdaBoost,’’ in Empirical Inference. Berlin, Germany: Springer, Oct. 2013, pp. 37–52.
[28] S. Taheri and M. Mammadov, ‘‘Learning the naive Bayes classifier with optimization models,’’ Int. J. Appl. Math. Comput. Sci., vol. 23, no. 4, pp. 787–795, Dec. 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY