Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/22216
DOI URL: http://dx.doi.org/10.21474/IJAR01/22216

RESEARCH ARTICLE

AN ENHANCED ASTUTE SYSTEM FOR PERSONALISED DIABETES DIAGNOSIS

Alekhya Palavelli[1], Dr. K.M. Rayudu[2] and Dr. M. Senthil[2]

1. PG Student.
2. Professor, Department of CSE, QIS College of Engineering and Technology (A), Ongole, Andhra Pradesh, India.

Manuscript Info

Manuscript History
Received: 04 September 2025
Final Accepted: 06 October 2025
Published: November 2025

Key words:-
Diabetes, Threat, Prediction, Machine Learning.

Abstract

Insulin resistance is the root cause of diabetes, a worldwide epidemic of chronic illness. Due to the lack of a cure for diabetes, the only method to lessen the threat of complications from the condition is via early diagnosis and treatment. Research towards the early diagnosis of diabetes using machine learning methods has been extensive. Improving prediction accuracy, however, is notoriously challenging due to the presence of missing values, irrelevant information, and uneven class distribution in the dataset. To better categorize the Pima Indians Diabetes Dataset (PIDD), we provide a Tree-Based machine learning approach in this study. To improve predictions, a Mutual Information (MI)-based feature selection approach is used to filter out irrelevant data. Finally, the Adaptive Boosting (AB) algorithm is used to boost the efficiency of Tree-Based algorithms. The Extra Tree (ET) technique used as the basis estimator of the AdaBoost classifier has the best accuracy (90.5% accuracy) when tested against experimental data. As a result, our suggested Tree-Based ML model may help doctors with diabetes diagnosis.

Introduction:-

A chronic ailment is a health problem that lasts longer than three months and has a significant negative impact on daily life for the patient. Chronic illnesses come in many forms, including cancer, Alzheimer's, arthritis, asthma, cardiovascular disease, and diabetes, to name a few. People who are afflicted with these conditions are often compelled to make drastic changes to their usual routine. Too much sugar in the blood causes diabetes. Insufficient insulin production by the pancreas or resistance to the effects of insulin generated leads to diabetes [1]. Insulin controls how much sugar the body absorbs from food. When blood sugar levels grow abnormally, it causes diabetes if it is not managed. There is a greater potential for permanent harm to systems including the nervous system and the cardiovascular system. Therefore, this illness heightens the danger of a number of potentially lethal disorders [1]. The World Health Organization classifies diabetes as one of three diseases: Insufficient insulin production in childhood results in Type I diabetes [2]. In most cases, it only manifests in young people [3]. When the body stops responding normally to insulin, it is diagnosed with Type II diabetes. Type II diabetes is more common in those who are middle-aged and older than 45 years old [3]. However, today it may also show up in younger people. The majority of people with diabetes today suffer from Type II [4]. High blood glucose levels cause gestational diabetes (Type III). Women who have never had diabetes before are often the ones who get the diagnosis during pregnancy.

By 2030 [5], the World Health Organization predicts that diabetes mellitus will rank as the sixth greatest cause of death worldwide. There will be 642 million young people with diabetes by 2040, according to a recent research [6]. About 1.6 million fatalities worldwide in 2016 were attributed to diabetes. This illness, however, cannot be eliminated. However, we have the ability to manage it by restricting blood glucose levels. The severity of diabetes may be reduced if caught early. This, however, is no simple endeavor. Information from the patient, including insulin levels, age, BMI, family history of diabetes, etc., is used to diagnose the condition [7]. Next, the doctor uses his or her expertise to make a call. However, the identification procedure may be lengthy and expensive. Due to the inexperience of the clinicians involved, it might potentially lead to erroneous diagnoses. Diabetic patients may be identified with help from computer-automated diagnostics.

Many studies have looked for ways to detect diabetes in its earliest stages. A number of medical databases have been created to hasten study in this area. Researchers have a difficult problem in properly classifying diabetes patients due to the non-linear, non-normal, and complicated character of the medical data. That's why we use a lot of ML and deep learning techniques to mine the dataset for insights and make diabetic illness predictions. Different machine learning techniques have been used to categorize people with diabetes. Most of them failed to adequately deal with outliers, missing data, and irrelevant characteristics, which led to poor classification accuracy. Preprocessing the data is crucial to achieving high classification accuracy while analyzing medical information. Despite these considerations, early identification of diabetes patients remains challenging due to the complexity of patient categorization [8].The focus of this research is on enhancing the accuracy of early diabetes diagnostic classifications. A machine learning-based Tree-Based prediction model for categorizing diabetes patients has been suggested. At first, we use the group average to fill in for missing data.After the missing data have been dealt with, the Robust scaling normalization method is used to find and deal with outliers.In order to solve the issue of uneven class distribution, an oversampling strategy is used. Then, a Tree-Based prediction model for diabetes early diagnosis is constructed using an efficient feature selection approach.

Relatedworks:-

Machine Learning in Healthcare:

General Application: Describe the role of machine learning in healthcare, particularly in disease prediction and diagnosis.Advantages of Machine Learning: Highlight how ML models can handle large datasets, recognize patterns, and provide predictive insights that may not be apparent through traditional methods.

Feature Selection and Engineering:

Importance of Features in Diabetes Prediction:Discuss how different features (e.g., age, BMI, blood pressure) impact the prediction accuracy.

Feature Selection Techniques:

Review common feature selection methods used in diabetes prediction models, such as recursive feature elimination and principal component analysis (PCA).

Datasets Used in Diabetes Prediction:

Pima Indians Diabetes Dataset:

Mention the widely-used Pima Indians Diabetes Dataset and how it's been applied in various studies.

Other Datasets:

Discuss other relevant datasets (e.g., electronic health records, patient survey data) used in developing diabetes prediction models.

Evaluation Metrics:

Common Metrics:

Explain the metrics commonly used to evaluate diabetes prediction models, such as accuracy, precision, recall, F1-score, and AUC-ROC.

Comparison of Results:

Provide a comparison of results from different models, noting which models perform best under various conditions.

Proposed System Architecture:-

PIDD was used in this study is shown in Table 1; it may be found in the UCI machine learning repository [9]. The median age of the samples in this data collection is 21 and the predominant gender is female. There are 768 samples here, and they are split evenly between two groups based on eight deterministic characteristics. Binary data is all that can be found in the class attribute. A score of '0' indicates a lack of diabetes (-ve), whereas a value of '1' indicates positive diabetes. There are 500 diabetes-negative samples and 268 diabetes-positive samples, or 65.1% and 34.9%, respectively. We have shown that medical data in the actual world are complicated, non-normal, and non-linear.Outliers, unreliable data, and missing values are commonplace. There are several elements that affect how well a classification algorithm performs; data preparation is one of them. Classification mistakes will be severe if data quality is not improved. Our study's ultimate goal is to improve the precision with which diabetes people can be predicted. For this reason, we have made every effort to efficiently manage the data. Missing value imputation, irrelevant feature removal, outlier identification, oversampling, and feature scaling are all examples of data preprocessing techniques. Fig.1 shows a flowchart of our suggested model. To scale features implies to normalize their values within a certain range [10].

This is a very important part of the preparation of data. The gap between a poor learner and a good one may be enormous if features are scaled before a machine learning model is created [11]. For many algorithms in machine learning and neural networks to converge more quickly and deal with outliers, feature scaling must sometimes be accelerated [12]. If we use the standard deviation and mean to normalize the data, it will be detrimental to any data that has too many outliers [13]. Our machine learning model would benefit greatly from the addition of features for dealing with outliers and scaling the data. Because of this, the data is scaled using Robust Scaler. The IQR has been utilized to normalize the data in this method [14]. The model is trained using k1 of the subsets, and then tested using the remaining subset. Model performance is determined by averaging the outcomes of the tests conducted on the k separate subsets, and this procedure is repeated k times throughout model construction. Ten-fold cross-validation was used to minimize error and variation in this study. The accuracy of a model's classifications is a common way to evaluate how well it performs as a prediction tool. Since the PIDD dataset has several issues, relying just on accuracy to evaluate the model's efficacy would be a mistake.This is why other statistical metrics have been consider ed for evaluating performance.

Table 1: Description of PIDD Dataset

S.N.	Name of Attribute	Attribute Description	Percentage o Missing Values
1	Pregnancies	No. of times pregnant	0
2	Glucose	Glucose concentration for 120 mins	0.65%
3	Blood Pressure	Diastolic blood pressure	4.55%
4	Skin Thickness	Skin fold thickness	29.55%
5	Insulin	Serum-insulin (2hours)	48.69%
6	BMI	Body mass index	0
7	Pedigree Function	Diabetes pedigree function	0
8	Age	Age in years	0

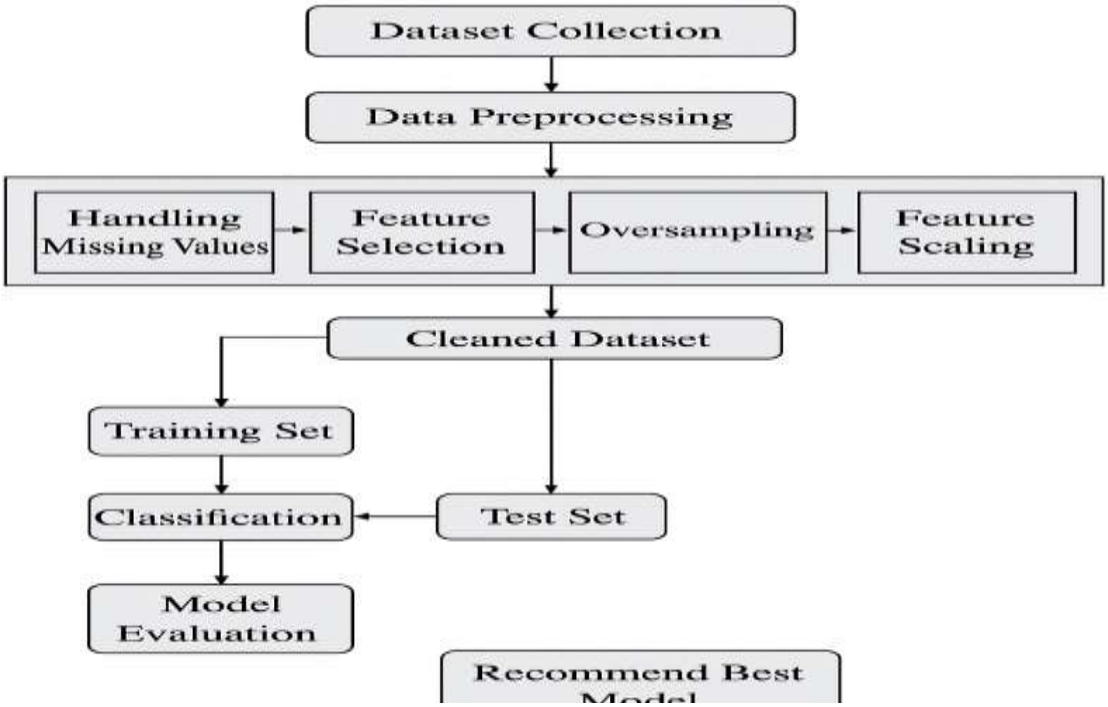

Fig.1 Proposed system architecture

Results:-

Our model is tested using the 10-fold cross-validation method for accuracy assessment. 80% of the data is used for training, 20% is used to verify its accuracy [15,16]. There are two stages to the categorization process. In the first stage, all characteristics are used in the classification process.Figure 5 shows that the accuracy of DT, RF, and ET are 82.8%, 86.1%, and 87.4%, respectively. These techniques are utilized as AdaBoost classifier's base estimator, which improves performance. Figure 4 shows that the accuracies of DT (83.5%), RF (86.4%), and ET (88.3%) are as follows. Figure 3 shows that there is an improvement in accuracy across all classifiers. AdaBoost combines many classifiers into one robust classifier, increasing performance across the board. Second, MI-FS is utilized to filter out irrelevant characteristics, such as BP and pedigree function. Figure 4 shows that the accuracies of DT (83.4%), RF (87.0%), and ET (88.6%) are as follows.Therefore, the assessment measures have improved due to the elimination of the superfluous characteristics. These techniques are then re-utilized as the AdaBoost classifier's base estimator, which leads to still another performance boost. As can be observed in Figure 5, the accuracy of DT, RF, and ET ranges from 84.25 percent to 87.2 percent to 90.5 percent.

Figure 5 shows that each successive classifier improves in accuracy.Free cloud service Google Colab [17] is utilized extensively for this project. Ten-fold cross-validation is used to evaluate our model's performance. Only 20% of the data is utilized to assess the model's correctness during training, whereas 80% is used for training [18]. Categorization happens in two phases. All of the criteria are considered during the initial phase of the categorization procedure. The figures in Figure 4 demonstrate that the reliability of DT is 82.8 percent, RF is 86.1 percent, and ET is 87.5 percent. AdaBoost uses these methods as its basis estimator, leading to significant gains in classification accuracy. Figure 4 displays the following accuracy rates for DT (83.5%), RF (86.4%), and ET (88.3%). In Figure 3, we can see that the accuracy of all classifiers has increased. AdaBoost is a method that improves performance by combining many classifiers into a single powerful classifier. Next, MI-FS is used to eliminate extraneous factors like BP and pedigree function. Fig. 4 displays the following accuracy rates for DT (83.4%), RF (87%) and ET (88.6%). As a result of removing the unnecessary factors, the evaluation criteria have become more accurate. The AdaBoost classifier takes use of these methods again by using them as its base estimator. Figure 5 shows that the DT, RF, and ET all have an accuracy of between 84.25 and 87.2%. The accuracy of the classifiers increases as seen in Figure 5.

Fig.2 Diabetes predictor

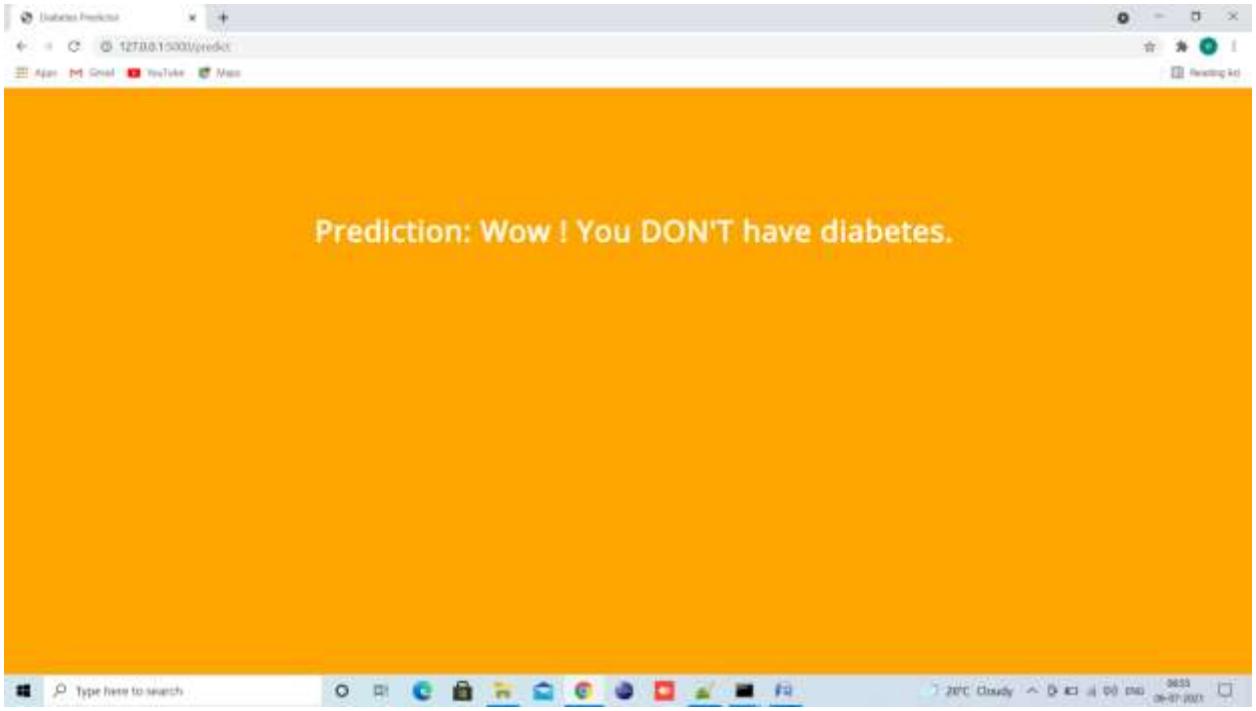

Fig.3 Diabetes prediction

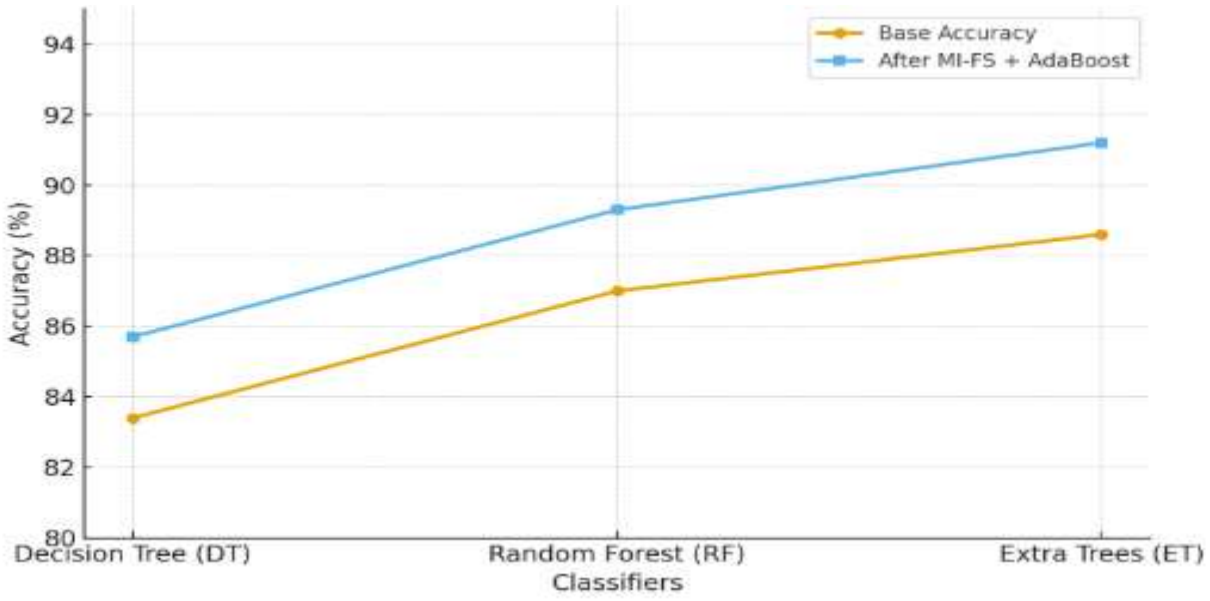

Fig.4 Accuracy Curves for Diabetes Prediction Classifiers

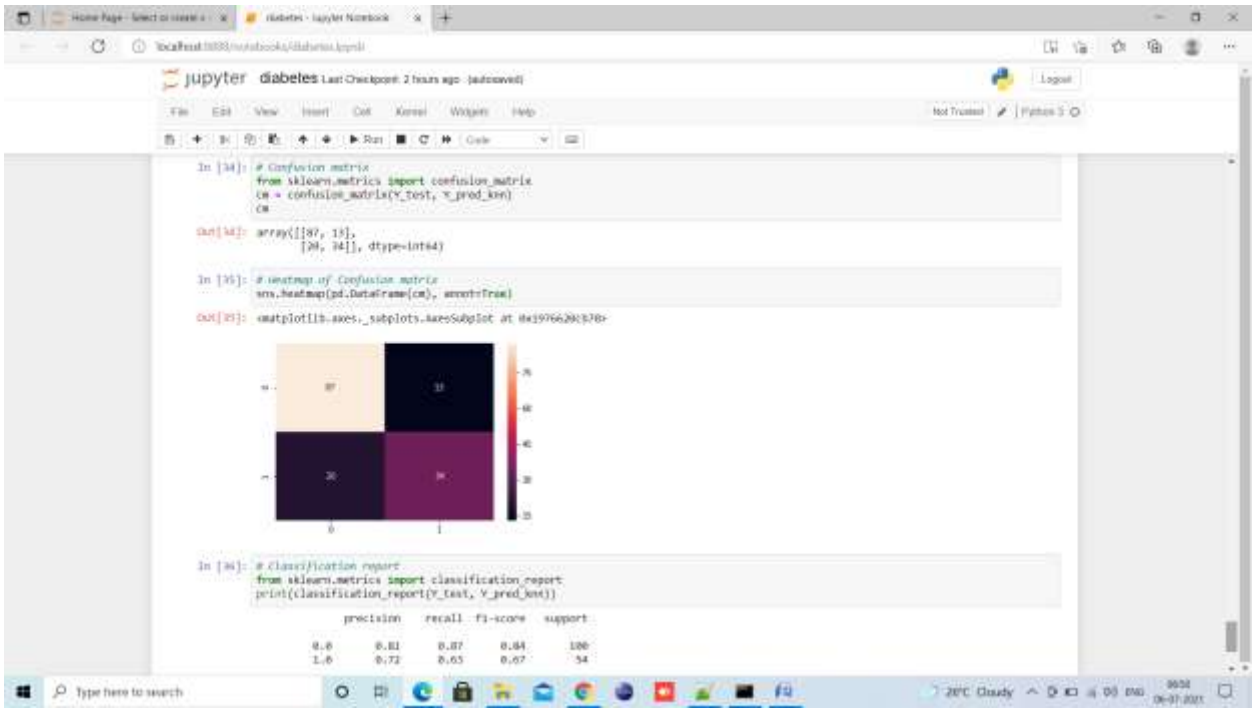

Fig.5 Confusion matrix

Conclusion:-

In this study, we compare and contrast different methods for categorizing people with diabetes. During data preparation, we oversampled the data and used mean imputation to deal with missing value and class distribution issues.Since then, a MI-based feature selection approach has been used to get the top-notch information. We have used Tree-Based machine learning techniques for data categorization.Also, AdaBoost classifier uses these methods as its foundational base estimator to achieve even higher levels of precision. The results of the comparison show that the Extra tree technique combined with the AdaBoost classifier achieves the best results. While three different tree-based classifiers are utilized, this study is limited since only MI is used to pick features. We want to experiment in

the future with a number of feature selection strategies, as well as a number of transfer learning and deep learning-based categorization strategies.

References:-

1. "Diabetes," [Online]. Available: https://www.who.int/newsroom/fact-sheets/detail/diabetes. [Accessed: 06-May-2020].
2. Maniruzzaman, M., Rahman, M.J., Al-Mehedi Hasan, M. et al.Accurate Diabetes Risk Stratification Using Machine Learning: Roleof Missing Value and Outliers. J Med Syst 42, 92, (2018).
3. "Diabetes Daily," [Online]. Available:https://www.diabetesdaily.com/learn-about-diabetes/what-isdiabetes/how-many people-have-diabetes/. [Accessed: 06-May-2020].
4. Ayon, Safial & Islam, Md. (2019). Diabetes Prediction: A DeepLearning Approach. International Journal of Information Engineeringand Electronic Business. 11. 21-27.
5. Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis,"DMP_MI: An Effective Diabetes Mellitus Classification Algorithmon Imbalanced Data With Missing Values," in IEEE Access, vol. 7,pp. 102232-102238, 2019.
6. Deepti Sisodia, Dilip Singh Sisodia, "Prediction of Diabetes usingClassification Algorithms," Procedia Computer Science, vol. 132, pp.1578-158, 2018.
7. Dr. S. Jafar Ali Ibrah,K M Rayudu,Setti Anitha ,K.Anitha,PArun Nambi "Brain Abnormality Detection and Analysis By Using MRI of Brain Through Multiple Sclerosis Lesions Detection And Analysis,"1st IEEE International conference on Innovations in High-Speed Communication and Signal Processing (IEEE-IHCSP)-March-2023.
8. H. Roopa and T. Asha, "A Linear Model Based on PrincipalComponent Analysis for Disease Prediction," in IEEE Access, vol. 7,pp. 105314-105318, 2019.
9. A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A DecisionSupport System for Diabetes Prediction Using Machine Learning andDeep Learning Techniques," 2019 1st International Informatics andSoftware Engineering Conference (UBMYK), Ankara, Turkey, 2019,pp. 1-4.
10. K Kannadasan, Damodar Reddy Edla, Venkatanareshbabu Kuppili,"Type 2 diabetes data classification using stacked autoencoders indeep neural networks," Clinical Epidemiology and Global Health,vol. 7, no. 4, pp. 530-535, 2019.
11. "UCI machine learning repository," [Online]. Available:https://archive.ics.uci.edu/ml/index.php. [Accessed: 06-May-2020].
12. A. I. Champa, M. F. Rabbi and N. Banik, "Improvement inHyperspectral Image Classification by Using Hybrid SubspaceDetection Technique," 2019 International Conference on SustainableTechnologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2019, pp. 15.
13. "Mutual information,"[Online]. Available:https://en.wikipedia.org/wiki/Mutual_information. [Accessed: 06-May-2020].
14. "ML | Feature Scaling - Part 2," [Online]. Available:https://www.geeksforgeeks.org/ml-feature-scaling-part-2/.[Accessed: 06-May-2020].
15. ["All about Feature Scaling," [Online]. Available:https://towardsdatascience.com/all-about-feature-scalingbcc0ad75cb35.[Accessed: 06-May-2020].
16. S. M. M. Hasan, M. A. Mamun, M. P. Uddin and M. A. Hossain,"Comparative Analysis of Classification Approaches for HeartDisease Prediction," 2018 International Conference on Computer,Communication, Chemical, Material and Electronic Engineering(IC4ME2), Rajshahi, 2018, pp. 1-4.
17. "Google Colaboratory," [Online]. Available:colab.research.google.com. [Accessed: 12-May-2020].
18. Mebrahtu, A., & Srinivasulu, B. (2017). Data mining techniques for building healthcare information system (HCIS). International Journal of Recent Engineering Research and Development, 2(8), 20–27.