Review of Computer Engineering Research
2022 Vol. 9, No. 2, pp. 71-82.
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v9i2.3037

HARD VOTING META CLASSIFIER FOR DISEASE DIAGNOSIS USING MEAN DECREASE IN IMPURITY FOR TREE MODELS

Ifra Altaf[1]
Muheet Ahmed Butt[2+]
Majid Zaman[3]

[1,2]*Department of Computer Sciences, University of Kashmir, Srinagar, J&K, India.*
[1]*Email: hussainifra3@gmail.com*
[2]*Email: ermuheet@gmail.com*
[3]*University of Kashmir, Srinagar, J&K, India.*
[3]*Email: zamanmajid@gmail.com*

(+ Corresponding author)

Article History
Received: 29 March 2022
Revised: 18 May 2022
Accepted: 2 June 2022
Published: 24 June 2022

Keywords
Voting ensemble
Bagging meta-estimator
Mean decrease in impurity
Tree models
Mean squared error.

ABSTRACT

To predict and detect various diseases, machine learning techniques are increasingly being used in the field of medical science. This study puts forward a bagging meta-estimator and feed forward neural network based voting ensemble with mean decrease in impurity feature selection to classify the disease datasets. The work was carried out using the Jupyter notebook data analysis tool, and Python 3 is used as a programming language. In this study, two chronic disease datasets - Indian Liver Patient dataset and the PIMA Indians diabetes dataset are used for building and testing the proposed model. The datasets are split into training and testing data in the ratio of 70:30. The experimental results illustrate that our proposed voting ensemble has an improved performance compared to the individual base learners. We also compared the accuracy of the model before and after the application of feature reduction technique. The results revealed that the accuracy increased with the removal of unimportant features. By using the proposed ensemble model, the average MSE, bias and variance were calculated as 0.311, 0.217 and 0.094 respectively for ILPD dataset. Similarly for PIMA dataset, the average MSE, bias and variance were calculated as 0.233, 0.186 and 0.047 respectively. These statistical parameters record a low score for ensemble classifier as compared to the individual constituent classifiers.

Contribution/Originality: The paper's primary contribution is finding a model that rationally treats the bias-variance trade-offs of a classification model thereby reducing its bias and variance. The proposed ensemble classifier achieves a good fit compared to the individual constituent classifiers.

1. INTRODUCTION

Artificial Intelligence (AI) is rapidly transforming our world. With time it has spread into many business areas [1-3] and is originating in medical field too with the increase in the complication and evolution of data in biological sciences, AI is gradually being applied within the field. The AI technologies such as machine learning (ML) and deep learning (DL) [4-7] have played a foremost role in the detection and prediction of diseases [8, 9] either by means of the disease symptom datasets [10, 11] or medical image datasets, which have helped the doctors in a positive way. Diagnosis of a disease is the critical step in today's disease control world. Machine learning algorithms have abundant potential to categorize the datasets [12, 13] of the medical domain and thus facilitate the diagnosis, prognosis, and development of treatment procedures and much more. The rudimentary task of machine learning is to create good models from the datasets [14] that best identify or forecast our required outcome. Instead of anticipating one model to be the most accurate predictor or detector [15, 16] an ensemble method comprising many models to produce one final detection or prediction model can be relied upon.

Motivated by the human collective decision making procedure, ensemble learning is also known as the combination of classifiers. It is a broad-spectrum meta-approach to machine learning [17-20] that pursues enhanced predictive performance [21] by joining the predictions from several models. The ML models that are included in ensemble learning are known as weak learners [22-26]. The final result of the ensemble model is achieved by grouping the result of its individual constituent weak learner [27, 28]. The concept of diversity is the main constituent of the realization of ensemble models [29].

The research study stressed mainly on creating a model that systematically treats the bias-variance trade-offs of a classification model thereby reducing its bias and variance. The study employed a hard voting ensemble that classified the data based upon the class labels and weights associated with each individual classifier taken. The novel voting based Meta-Estimator and Feed Forward Neural Network tries to reduce variance and bias by employing the bagging and boosting techniques. Further, the novel ensemble is used with GINI Importance or Mean Decrease in Impurity feature importance measures for tree models to achieve better classification performance based on accuracy. The proposed classifier takes the integration of the predictions from the Extreme Gradient Boosting or XGBoost (XGB), Bagging Meta Estimator (BE) and the Multilayer Perceptron (MLP). The formula for the functions of these algorithms is given in Equation1, Equation 2 and Equation 3.

$$obj(\theta) = \sum_{i}^{n} l(y_i - y_{\hat{\imath}}) + \sum_{j=1}^{j} \Omega(p_j) \tag{1}$$

Where, p_j means a prediction that comes from the jth tree.

$$h(x) = \phi(x) = s(b(1) + W(1)x) \tag{2}$$

$$o(x) = \phi(x) = G\big(b(2) + W(2)h(x)\big) \tag{3}$$

Where, b(1), b(2) are bias vectors, W(1), W(2) are the weight matrices ,G and s are the activation functions.

The voting ensemble classifier is a powerful meta-classifier that measures the weaknesses of individual constituent models on particular datasets. The implementation was performed using the Python programming language and Jupyter notebook data mining tool. Several performance measures were used to measure the performance of the proposed ensemble model. The rest of the paper is divided into the following sections: Section 2 discusses the literature review; Section 3 provides the proposed work by explaining the dataset and experimental setup. ; Section 4 gives the results analysis and discussion. Section 5 concludes the paper and suggests future work.

2. LITERATURE REVIEW

By borrowing the machine learning approaches – individual as well as ensemble techniques, many publications have reported their methods to classify the liver and diabetic patients by using the Indian Liver Patient Dataset (ILPD) and Pima Indian Diabetes Dataset (PIMA). The analysis done by some authors to predict or detect the liver and diabetes diseases with the help of different machine learning algorithms are given below:

Kabir and Simone [30] proposed the super learning, also known as the stacked-ensemble, to find the ideal weighted average of different learning models. The model performed better than the individual base learners. The super learner comprised gradient boosting machine (GBM), deep neural network (DNN) and random forest (RF). With three base models, the stacked ensemble gave the accuracy value of 73.4% for ILPD Kabir and Simone [30]. Bihter [31] used the high performance neural network on ILPD that gave the training accuracy of 71.95% and validation accuracy as 73.28% Bihter [31]. Abedini, et al. [32] put forward an ensemble model comprising artificial neural network (ANN), logistic regression (LR) and decision tree (DT) to classify the PIMA dataset. The accuracy obtained by the authors was calculated as 83.08% Abedini, et al. [32]. Razali, et al. [33] proposed the use of neural network and bayesian model on ILPD and got the accuracy result of 66.85% and 70.52% respectively Razali, et al. [33]. Barik [34] used a hybrid machine learning technique XGBoost method on PIMA dataset and got the prediction value 74.10% Barik [34]. Singh, et al. [35] applied the variants of support vector machine (SVM) on the ILPD dataset and found that the Coarse Gaussian SVM achieved the highest accuracy of 71.4% in predicting the disease [35].

3. PROPOSED WORK

This research study involved a binary classification problem and proposes a Majority Voting Ensemble model. The model is a Meta-Estimator and Feed Forward Neural Network based voting ensemble trained with MDI feature selection to classify the disease datasets. The model takes in the instances of a number of patients whose records were acquired from the online repository. The raw datasets acquired from the online repository were pre-processed to get clean and noise free data. MDI feature importance measure for random forests was used to extract the important features from the datasets in to ensure better accuracy based on the given data. The supervised learning algorithms used to make the ensemble were XGB, BE and MLP. The prediction outputs of the individual classifiers were combined in voting ensemble approach. The average Mean Squared Error (MSE) was used to calculate the average squared difference between the estimated values and the actual value of the proposed model. The values obtained were compared with that of the individual classifiers.

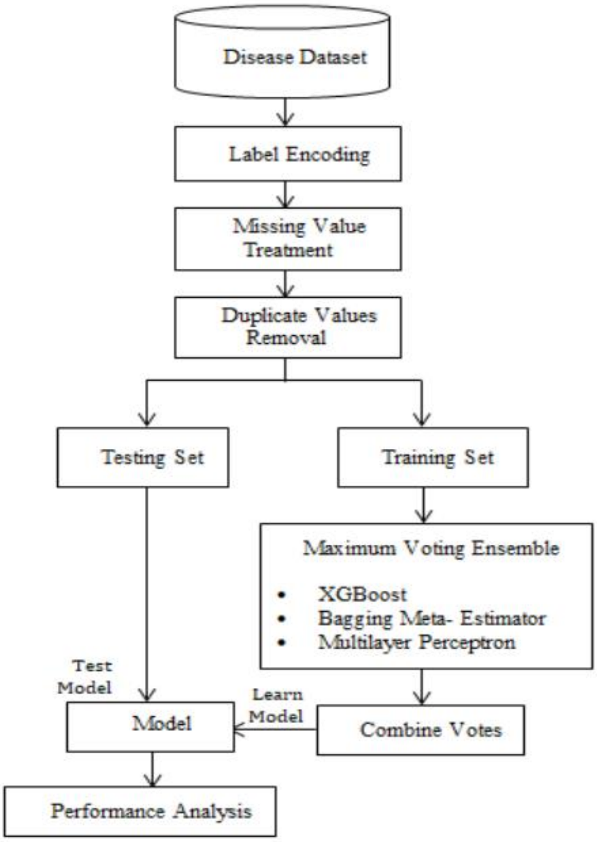


Figure 1. Working diagram for the proposed model design.

In this study, two chronic diseases, ILPD and PIMA, were used for building and testing the proposed model. The performance of the model is given in terms of accuracy, precision, recall, support and f1-scores which help to validate the results of our proposed ensemble. The Python 3.7 programming language was used to implement the proposed method. Figure 1 shows the flow diagram demonstrating the workflow diagram of the proposed research study.

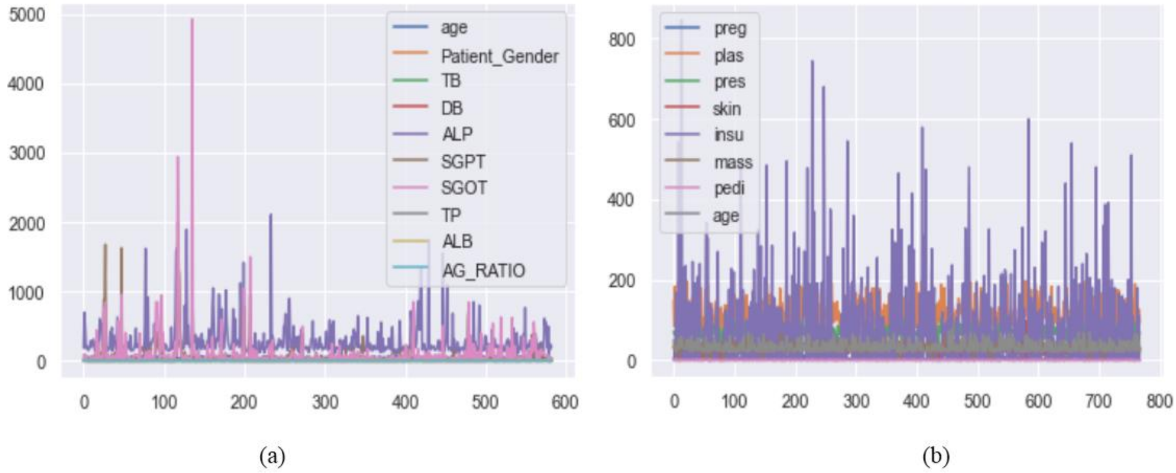
A. *Objectives*

The main objectives of this research study were to:

- Extract features from disease dataset using the feature importance with a forest of trees based on mean decrease in impurity.
- Implement Majority Voting Ensemble classifier to classify the various diseases.
- Compare the accuracy of the model before and after the application of feature reduction technique.
- Compare the accuracies of ensemble model with its individual base learners.
- Calculate the MSE, average bias and average variance of the individual classifiers as well as on the proposed model.
- Compare the accuracies with other experiments from previous literature to prove the significance of the proposed model.

3.1. Dataset

The disease datasets that were used in this research study were acquired from Kaggle data science community. The datasets consist of the Indian Liver Patient Dataset (ILPD) [36] and the PIMA Indians diabetes dataset [37]. The ILPD dataset has 583 records and 10 attributes besides a selector-field that classifies liver patients. It consists of 416 liver disease patients and 167 non-liver disease patients. The PIMA dataset has 768 records and 8 attributes besides a selector-field that classifies diabetic patients. It consists of 500 diabetic patients and 268 non-diabetic patients. Figure 2 shows the features of ILPD and PIMA respectively. Figure 3 gives the pair plot of the disease datasets taken. The figures illustrate the matrix of relationships between each attribute in the dataset. It visualizes the relationships between each attribute present in the ILPD and PIMA datasets respectively, which helped with the immediate analysis of the data.



(a) (b)

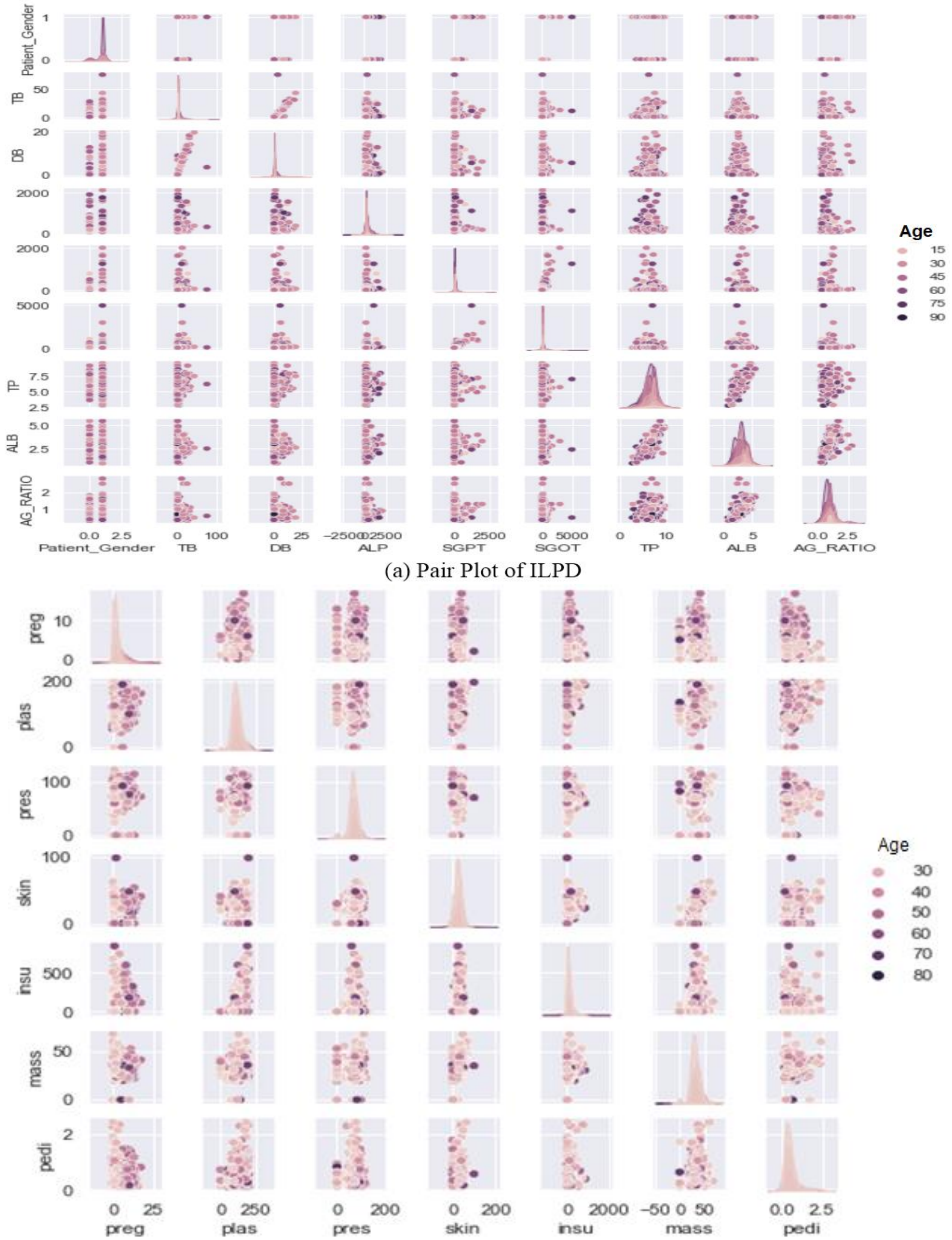
Figure 2. Distribution of the features of disease datasets (a) ILPD and (b) PIMA.

3.2. Experimental Setup

B. *Data Preprocessing*

The data preprocessing being one of the data mining strategies was used to deal with the missing values, duplicate values as well as feature selection. Since sklearn does not allow non-numerical data, we therefore transformed categorical data into dummy variables using the LabelEncoder function.

Next, the missing values of the dataset were mitigated. Since the maximum features of the acquired disease datasets follow a skewed distribution, the missing values of the collected dataset were, therefore, handled with the SimpleImputer class of scikit-learn library. The missing values were directly replaced with the 'median' to maintain the important information of the dataset. The duplicates in the datasets were subsequently removed and the datasets were then subjected to feature reduction technique.



(a) Pair Plot of ILPD

(b) Pair Plot of PIMA

Figure 3. Pair plot of disease datasets pertaining to patient age. (a) (b)

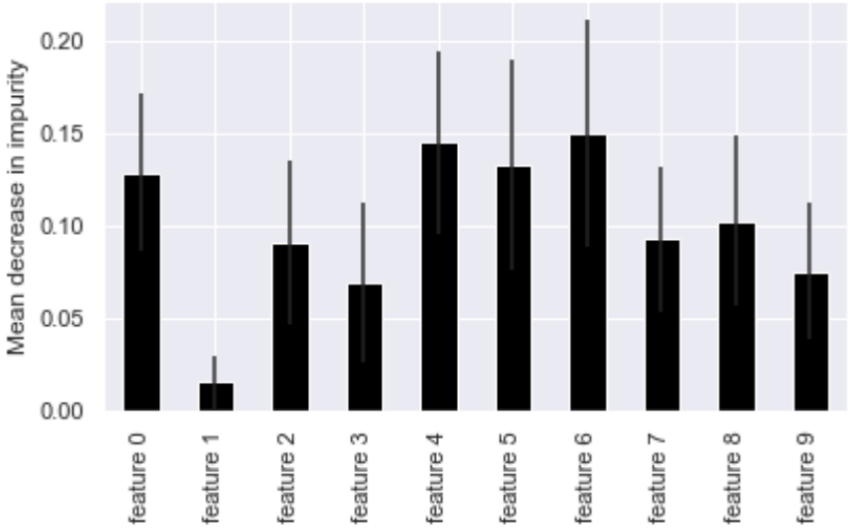


Figure 4. Significant features using MDI on ILPD.

The features of the acquired datasets were toned down by using forest of trees to assess the importance of features. We used GINI Importance, also known as Mean Decrease in Impurity (MDI), to measure feature importance with Random Forest model. The mean and standard deviation of accumulation of the impurity decrease within each tree was calculated using the fitted attributes of the disease datasets. Seven features and five features were selected from ILPD and PIMA respectively Figure 4 and Figure 5.

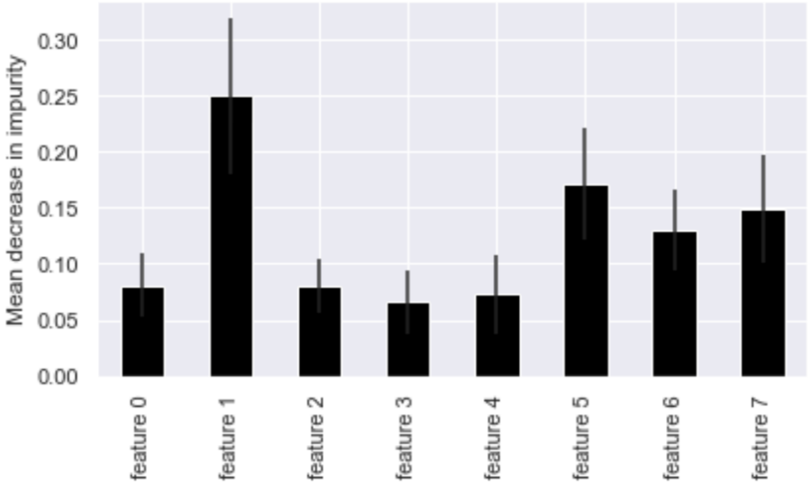


Figure 5. Significant features using MDI on PIMA.

Within each tree, the calculation of mean and standard deviation of accumulation of the impurity decrease contributes towards the feature importance based on mean decrease in impurity. Table 1 gives the time taken to compute the feature importance of the selected two datasets.

Table 1. Elapsed time to compute the feature importance using MDI.

Dataset	Elapsed Time
ILPD	0.04 seconds
PIMA	0.03 seconds

C. Model Training

The algorithms were implemented using the Python programming language. The datasets were split into training and testing data in the ratio of 70:30 using the 'train_test_split' class of sklearn machine learning library.

The proposed Majority Voting Ensemble, as well as individual base classifiers, was built with the training data while the performance was evaluated on the basis of the test data Figure 6. The research work was carried out in a free online cloud based Jupyter notebook environment - Google Colab. The proposed ensemble technique comprising of XGBoost, Bagging Meta-Estimator and Multilayer Perceptron (both individually as well as in the ensemble form) was implemented on the full datasets and then on reduced feature sets. Table 2 shows the pseudo code for Hard Voting Meta Classifier for disease diagnosis and prediction based on mean decrease in impurity (feature selection) for tree models.

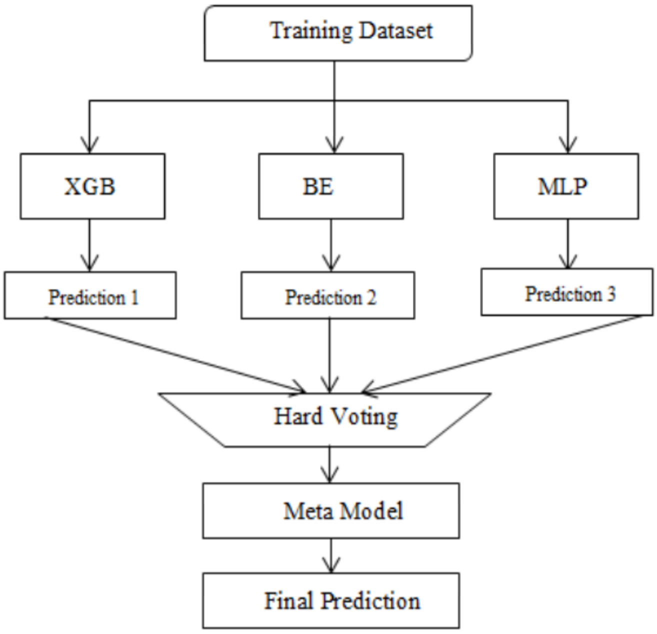


Figure 6. The proposed hard voting ensemble.

Table 2. Pseudo-code for hard voting meta classifier algorithm.

```
Disease_dataset = ILPD.csv, PIMA.csv
Procedure: SPLIT_DATA (disease_dataset)
train_data, test_data <- train_test_split(diabetes_attributes) return train_data, test_data
model_estimators <- [ ]
model_estimators.append(XGBClassifier)
model_estimators.append(BaggingClassifier)
model_estimators.append(MLPClassifier)
hard_voting <-VotingClassifier(estimators <-model_estimators)
hard_voting.fit(train_data, test_data)
procedure: Voting Classifier (hard_voting, train_data, train_label, test_data, test_label)
testing_diagnosis <-hard_voting.diagnose(test_data)
training_diagnosis <- hard_voting.diagnose(train_data).
```

D. Results

The objective of this study was to find out if a patient will acquire a particular disease or not. Firstly, from our proposed method we trained and tested the datasets with individual learners. We then trained and tested the datasets with their voting ensemble. Later, we prioritized the data by using the MDI based feature selection technique and removed the attributes that were less important in each dataset. After the implementation of the proposed ensemble model, Table 3 and Table 4 present the classification report and confusion matrix of ILPD and PIMA respectively.

Table 3. Confusion matrix and classification report of ILPD.

True Negative	False Negative	True Positive	False Positive		
113	9	15	37		
Accuracy	73.5%				
Classification Report					
	1	2	Accuracy	Macro Avg	Weighted Avg
Precision	0.75	0.62	0.73	0.68	0.71
Recall	0.92	0.28	0.73	0.60	0.73
f1-Score	0.83	0.39	0.73	0.61	0.70
Support	122	52.0	0.73	174	174

Table 4. Confusion matrix and classification report of PIMA.

True Negative	False Negative	True Positive	False Positive		
137	20	44	30		
Accuracy	78.3%				
Classification Report					
	1	2	Accuracy	Macro Avg	Weighted Avg
Precision	0.82	0.68	0.78	0.75	0.77
Recall	0.87	0.59	0.78	0.73	0.78
f1-Score	0.84	0.63	0.78	0.74	0.77
Support	157	74.0	0.78	231	231

Table 5 gives the accuracy comparison of the proposed hard voting based ensemble learner with its constituent classifiers. Table 6 gives the average MSE, bias and variance indicators of each individual classifier as well as their majority voted ensemble. The proposed ensemble classifier achieved a good fit compared to the individual constituent classifiers. By looking at the performance of the model, we observed that the mean square error of the classifier starts to surge upon changing either the classifiers or their parameters.

Table 5. Accuracy comparison between base classifiers and voting ensemble.

	XG Boost	Bagging Meta Estimator	MLP	Voting Ensemble
Without Feature Reduction				
ILPD	63.22%	63.22%	59.20%	65.52%
PIMA	75.76%	73.59%	62.34%	74.03%
With MDI Feature Reduction				
ILPD	70.69%	70.11%	70.11%	73.56%
PIMA	70.56%	71.86%	67.53%	78.35%

Table 6. Mean squared error, bias and variance statistics.

Dataset		XG Boost	Bagging Meta Estimator	MLP	Voting Ensemble
Without Feature Reduction					
ILPD	Average MSE	0.34	0.33	0.36	0.32
	Average Bias	0.23	0.23	0.19	0.23
	Average Variance	0.10	0.09	0.17	0.09
PIMA	Average MSE	0.25	0.25	0.34	0.23
	Average Bias	0.16	0.16	0.23	0.18
	Average Variance	0.08	0.08	0.11	0.04
With MDI Feature Reduction					
ILPD	Average MSE	0.33	0.32	0.32	0.31
	Average Bias	0.21	0.22	0.22	0.21
	Average Variance	0.11	0.09	0.09	0.09
PIMA	Average MSE	0.25	0.25	0.32	0.23
	Average Bias	0.16	0.17	0.23	0.18
	Average Variance	0.08	0.07	0.09	0.04

4. RESULTS ANALYSIS AND DISCUSSION

By using the maximum voting with feature selection strategy, we achieved a greater accuracy in classifying the diseases. Figure 7 gives the comparison between accuracies attained by proposed Voting ensemble before and after using the feature reduction technique. Figure 8 gives the error statistics of the individual as well as ensemble classifier.

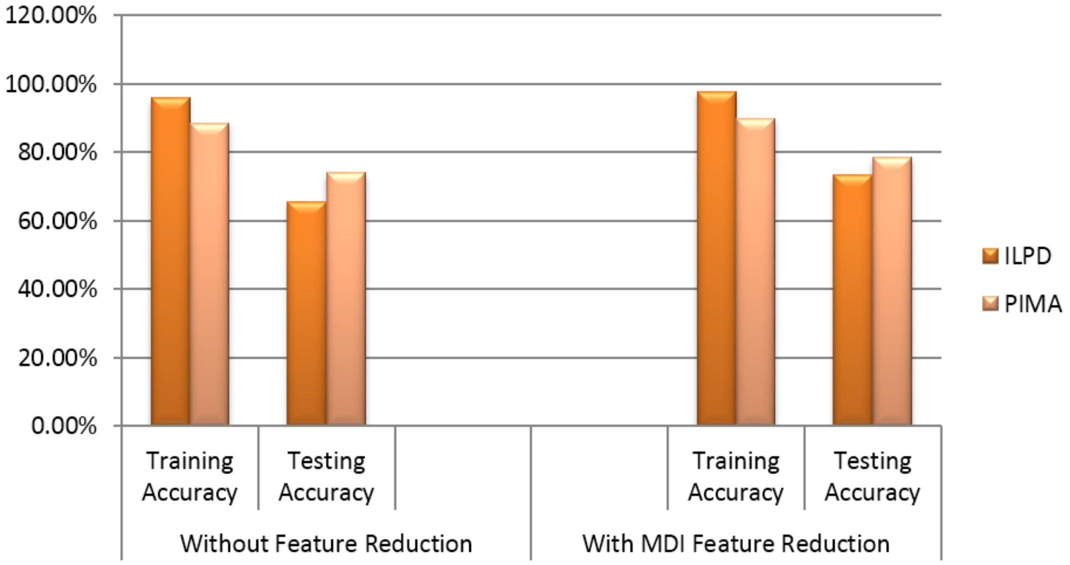


Figure 7. Accuracy comparison of voting ensemble with and without feature reduction.

The proposed ensemble algorithm was implemented on the standard datasets and the overall performance was calculated. It was observed that the ensemble method outperformed the other individual classifiers in terms of accuracy, MSE, bias and variance. The mean square error of the proposed ensemble varies between 0.233% and 0.311%.

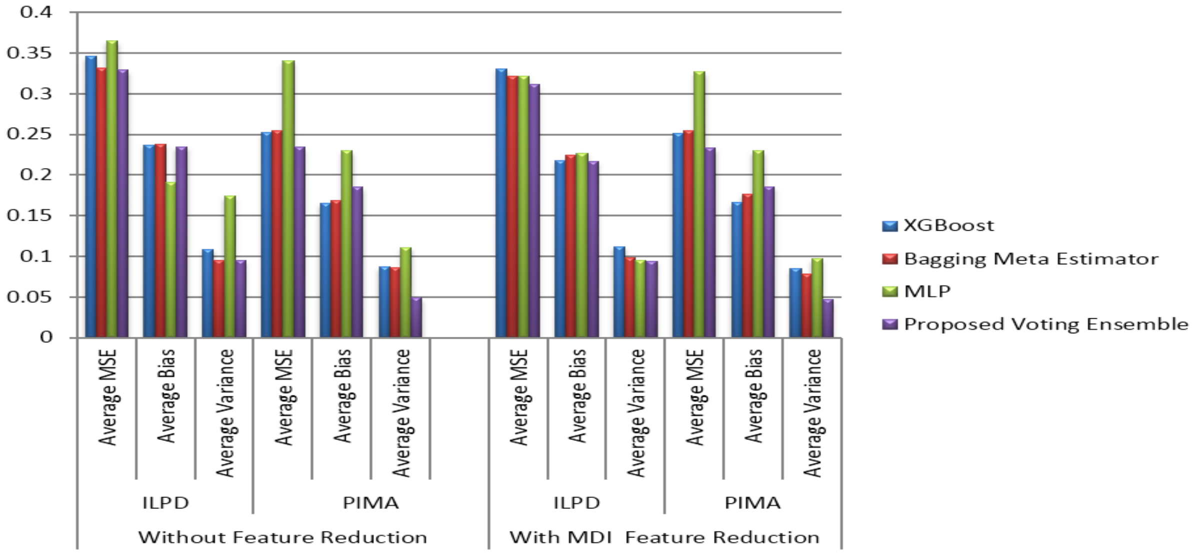


Figure 8. Mean squared error statistics for disease datasets using individual and ensemble classifier.

This range delineates the effective performance of the proposed ensemble model. The accuracy of bagging and boosting algorithms taken separately was less than the accuracy obtained when their predictions or outputs were combined. It was also observed that the MSE, bias and variance of the combined ensemble model was slightly lower than that of the individual classifiers. The ensemble works by reducing the variance and bias by combining the bagging and boosting techniques. The feature reduction played a dynamic role in selecting the most suitable features for classifying the disease. The shortcomings of boosting and bagging techniques were moderated by their unification. Table 6 demonstrates the assessment among various studies done on disease classification in recent years with the projected model of the paper. Furthermore, Figure 8 presents the error statistics of the proposed model for the two datasets while using and without using the feature reduction technique. The results show that the error statistics tend to be lower when the feature reduction technique is applied.

5. CONCLUSION AND FUTURE WORK

The maximum voting ensemble algorithm proposed in this paper assists in detecting and predicting liver as well as diabetic patients from certain datasets. The algorithms were implemented using the Python programming language. The prediction or output of different base learners was been combined using the maximum voting technique to improve prediction rate. The experimental results show that the hybrid ensemble machine learning model along with the MDI feature reduction technique gave the best accuracy. Also, the MSE of the proposed ensemble along with the bias and variance reduce marginally in the case of the ensemble technique. MDI is an impurity-based feature importance technique and for high cardinality features it can, at times, be ambiguous. Permutation feature importance, also known as the Mean Decrease in Accuracy (MDA), can be used in future to see the influence on the accuracy results. The prediction score depends mainly on the type of the data used to carry out the experiments. We cannot reach a generalized inference or arrange an inception mechanism to summarize that the proposed model achieves better prediction and detection accuracy. For that there is a need for rigorous analysis that takes a larger number of datasets in to consideration. In the future, the proposed model will be used to classify other diseases such as Parkinson's, Cancer, Alzheimer's, etc. and infer whether the algorithm is competent or not. Table 7 presents a comparison of our proposed algorithm with the existing approaches for disease classification in terms of accuracy.

Table 7. Comparison of result with existing system.

Author	Year	Model	Dataset	Accuracy
Kabir and Simone [30]	2019	Stacked Ensemble	ILPD	73.4%
Bihter [31]	2020	Neural Network	ILPD	73.28%
Razali, et al. [33]	2020	Bayesian Model	ILPD	70.52%
Barik [34]	2021	Hybrid XGBoost	PIMA	74.10%
Singh, et al. [35]	2021	Coarse Gaussian SVM	ILPD	71.4%
Altaf, et al. [8]	2022	Voting Ensemble with MDI	PIMA	78.35%
Altaf, et al. [8]	2022	Voting Ensemble with MDI	ILPD	74.03%

Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study.

REFERENCES

[1] S. A. Fayaz, Z. Majid, and A. B. Muheet, "Knowledge discovery in geographical sciences—A systematic survey of various machine learning algorithms for rainfall prediction," in *International Conference on Innovative Computing and Communications. Springer, Singapore*, 2022.

[2] M. Ashraf, M. A. Syed, A. G. Nazir, A. S. Riaz, Z. Majid, A. K. Sameer, and A. S. Aftab, "Prediction of cardiovascular disease through cutting-edge deep learning technologies: An empirical study based on tensorflow, pytorch and keras," in *In International Conference on Innovative Computing and Communications, pp. 239-255. Springer, Singapore*, 2020.

[3] R. Mohd, A. B. Muheet, and Z. B. Majid, "Grey wolf-based linear regression model for rainfall prediction," *International Journal of Information Technologies and Systems Approach*, vol. 15, pp. 1-18, 2022.

[4] Z. Majid, K. Sameer, and A. Muheet, "Analytical comparison between the information gain and gini index using historical geographical data," *International Journal of Advanced Computer Science and Applications*, vol. 11, pp. 429-440, 2020.

[5] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system for educational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471-1483, 2020.Available at: https://doi.org/10.1016/j.procs.2020.03.358.

[6] R. Mohd, M. A. Butt., and M. Z. Baba., "SALM-NARX: Self adaptive LM-based NARX model for the prediction of rainfall," in *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on. IEEE*, 2018, pp. 580-585.

[7] M. Ashraf, Z. Majid, and A. Muheet, "To ameliorate classification accuracy using ensemble vote approach and base classifiers." In emerging technologies in data mining and information security," ed Singapore: Springer, 2019, pp. 321-334.

[8] I. Altaf, A. B. Muheet, and Z. Majid, "Disease detection and prediction using the liver function test data: A review of machine learning algorithms," in *International Conference on Innovative Computing and Commu-nications. Springer, Singapore*, 2022.

[9] M. Ashraf, Z. Majid, and A. Muheet, "Performance analysis and different subject combinations: An empirical and analytical discourse of educational data mining," in *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE*, 2018, pp. 287-292.

[10] I. Altaf, A. B. Muheet, and Z. Majid, "A pragmatic comparison of supervised machine learning classifiers for disease diagnosis," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE*, 2021.

[11] M. Ashraf, M. Zaman, and M. Ahmed, "Using ensemble StackingC method and base classifiers to ameliorate prediction accuracy of pedagogical data," *Procedia Computer Science*, vol. 132, pp. 1021-1040, 2018.Available at: https://doi.org/10.1016/j.procs.2018.05.018.

[12] S. A. Fayaz, Z. Majid, and A. B. Muheet, "Performance evaluation of GINI index and information gain criteria on geographical data: An empirical study based on JAVA and python," in *International Conference on Innovative Computing and Communications. Springer, Singapore*, 2022.

[13] N. M. Mir, S. Khan, M. A. Butt, and M. Zaman, "An experimental evaluation of bayesian classifiers applied to intrusion detection," *Indian Journal of Science and Technology*, vol. 9, pp. 1-7, 2016.Available at: https://doi.org/10.17485/ijst/2016/v9i12/86291.

[14] S. A. Fayaz, Z. Majid, and A. B. Muheet, "Numerical and experimental investigation of meteorological data using adaptive linear M5 model tree for the prediction of rainfall," *Review of Computer Engineering Research*, vol. 9, pp. 1-12, 2022.

[15] Z. Majid and A. B. Muheet, "Information translation: A practitioners approach," in *World Congress on Engineering and Computer Science (WCECS), San Francisco, USA*, 2012.

[16] S. A. Fayaz, M. Zaman, and M. A. Butt, "To ameliorate classification accuracy using ensemble distributed decision tree (DDT) vote approach: An empirical discourse of geographical data mining," *Procedia Computer Science*, vol. 184, pp. 935-940, 2021.Available at: https://doi.org/10.1016/j.procs.2021.03.116.

[17] H. M. Waseem, S. Jamwal, and M. Zaman, "Congestion control techniques in a computer network: A survey," *International Journal of Computer Applications*, vol. 111, pp. 7-10, 2015.Available at: https://doi.org/10.5120/19508-1112.

[18] S. A. Fayaz, I. Altaf, A. N. Khan, and Z. H. Wani, "A possible solution to grid security issue using authentication: An overview," *Journal of Web Engineering & Technology*, vol. 5, pp. 10-14, 2019.

[19] E. M. A. Butt, S. M. K. Quadri, and E. M. Zaman, "Star schema implementation for automation of examination records," in *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer*

Engineering (FECS). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[20] S. J. Sidiq, M. Zaman, and M. Ahmed, "How machine learning is redefining geographical science: A review of literature," *International Journal of Emerging Technologies and Innovative Research*, vol. 6, pp. 1731-1746, 2019.

[21] Ifra Altaf, A. B. Muheet, S. Majid Zaman, and J. Sidiq, "A Comparative Study of various data mining algorithms for effective liver disease diagnosis A decade review from 2010 to 201," vol. 6, pp. 980-995, 2019.

[22] M. Zaman, S. Quadri, and M. A. Butt, "Generic search optimization for heterogeneous data sources," *International Journal of Computer Applications*, vol. 44, pp. 14-7, 2012.Available at: https://doi.org/10.5120/6258-8404.

[23] M. Zaman and M. A. Butt, "Enterprise data backup & recovery: A generic approach," *International Organization of Scientific Research Journal of Engineering (IOSRJEN)*, pp. 2278-4721, 2013.

[24] S. Kaul, A. F. Sheikh, Z. Majid, and A. B. Muheet, "Is decision tree obsolete in its original form? A burning debate," *Artificial Intelligence Review*, vol. 36, pp. 105-113, 2022.

[25] M. Hassan, M. A. Butt, and M. Z. Baba, "Logistic regression versus neural networks: The best accuracy in prediction of diabetes disease," *Asi. J. of Comp. Sci. and Tech*, vol. 6, pp. 33-42, 2017.

[26] D. Nayak and E. M. A. Butt, "Empowering cloud security through sla," *Journal of Global Research in Computer Science*, vol. 4, pp. 30-33, 2013.

[27] V. Maheshwari, M. R. Mahmood, S. Sravanthi, N. Arivazhagan, A. ParimalaGandhi, K. Srihari, and V. P. Sundramurthy, "Nanotechnology-based sensitive biosensors for COVID-19 prediction using fuzzy logic control," *Journal of Nanomaterials*, pp. 1-8, 2021.Available at: https://doi.org/10.1155/2021/3383146.

[28] A. A. Shehloo, A. B. Muheet, and Z. Majid, "Factors affecting cloud data-center efficiency: A scheduling algorithm-based analysis," *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, p. 1136, 2021.

[29] C. Qi and X. Tang, "A hybrid ensemble method for improved prediction of slope stability," *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 42, pp. 1823-1839, 2018.Available at: https://doi.org/10.1002/nag.2834.

[30] M. F. Kabir and A. L. Simone, "Enhancing the performance of classification using super learning," *Data-Enabled Discovery and Applications*, vol. 3, pp. 1-13, 2019.Available at: https://doi.org/10.1007/s41688-019-0030-0.

[31] D. A. Ş. Bihter, "A comparative study on the performance of classification algorithms for effective diagnosis of liver diseases," *Sakarya University Journal of Computer and Information Sciences*, vol. 3, pp. 366-375, 2020.

[32] M. Abedini, A. Bijari, and T. Banirostam, "Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network," *Intern JAdvan Res Comp Commun Engin*, vol. 9, pp. 1-5, 2020.Available at: https://doi.org/10.17148/ijarcce.2020.9701.

[33] N. Razali, A. Mustapha, M. H. Abd Wahab, S. A. Mostafa, and S. K. Rostam, "A data mining approach to prediction of liver diseases," in *Journal of Physics: Conference Series*, 2020, p. 032002.

[34] S. Barik, et al., "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. Intelligent and Cloud Computing," ed Singapore: Springer, 2021, pp. 399-409.

[35] L. Singh, R. J. Rekh, and P. S. Satya, "Classification of hepatic disease using machine learning algorithms. Advances in Biomedical Engineering and Technology," ed Singapore: Springer, 2021, pp. 161-173.

[36] A. Soni, "Performance analysis of classification algorithms on liver disease detection," presented at the 2021 IEEE Mysore Sub Section International Conference (MysuruCon), pp. 1-5. IEEE, 2021.

[37] R. Rousyati, A. N. Rais, E. Rahmawati, and R. F. Amir, "Pima Indians diabetes prediction database with adaboost and bagging ensemble," *EVOLUTION: Journal of Science and Management*, vol. 9, pp. 36-42, 2021.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.