Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MEDPOLY: A Proposed Approach for Cardiovascular Disease Prediction Using Various Ensemble Models

Sreekumari Sreelesh
Lovely Professional University Punjab
Rajni Bhalla
Lovely Professional University Punjab
Gursharan Singh
gursharan.dhot@gmail.com
Lovely Professional University Punjab

Research Article

Keywords: Ensemble model, Bagging, Boosting, Decision Tree, Neural net- works

Posted Date: September 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7081120/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MEDPOLY: A Proposed Approach for Cardiovascular Disease Prediction Using Various Ensemble Models

Sreekumari Sreelesh
sreekumari.sreelesh@gmail.com
Lovely Professional University
Punjab, India

Rajni Bhalla
rajni.b27@gmail.com
Lovely Professional University
Punjab, India

Gursharan Singh
gursharan.dhot@gmail.com
Lovely Professional University
Punjab, India

Corresponding Author: Gursharan Singh <gursharan.dhot@gmail.com

Abstract An ensemble model is a combination of different numbers of models for making predictions. Each model is trained separately and all models participate to solve the same issue. The output from all models is combined to make predictions. Research shows that the ensemble model provides better results as compared to individual models. In the 1990's, four popular machine learning models were introduced bagging, boosting, stacking, and voting to improve the performance of the ensemble model. The two-tailed test indicates ensemble model is best as compared to the single method. In this study, two ensemble models have been proposed that are applied to different datasets. The first model is a combination of decision tree, neural network, and logistic regression(DNL) that is applied to 4 different datasets. The second model name MEDPOLY is proposed as a combination of 24 different classifiers and tested using bagging, boosting, stacking, and voting classifiers. DNL ensemble model using stacking performs better as compared with bagging and provides an accuracy of 98% in predicting cardiovascular disease. MEDPOLY ensemble model that consists of 24 different classifiers is compared with DNL also. The result reveals that the DNL outperforms the MEDPOLY framework. Analysis indicates that DNL is capable of handling linear relationships and complex patterns and is also able to handle situations of overfitting. MEDPOLY when applied to a dataset that is not large or diverse enough, can lead to over-fitting. A combination of a smaller set of classifiers provides better and interpretable results.

Keywords: Ensemble model, Bagging, Boosting, Decision Tree, Neural networks

1 Introduction

Most researchers combined the multiple models into a single model to achieve good accuracy [8] [9] [29] [31]. It has been proved from the result that a combined model has better accuracy than a single classifier[18][23]. There are two important issues to take care of when choosing a classifier. First, each classifier must be accurate, meaning it should perform well while making predictions. Another benefit is overcoming errors of one classifier. If one classifier produces an output on a particular type of data, another classifier in an ensemble model must resolve that error. This makes the ensemble model more robust and efficient [19][28][27]. An Ensemble consists of two popular techniques one is known as Bagging and the other is known as Boosting.

Many ensemble methods like bagging and boosting have been created and tested for prediction. Ensemble techniques demonstrated that those techniques are very effective with decision trees[5][10][7][12][30] and very little work is done with neural networks. Fast training speed and well-established parameter settings lead to the use of decision trees. While working with the neural network, difficulty comes with processing time and selecting training parameters. We benefit from adding neural networks while making predictions for different datasets. Neural networks as individuals provide more accurate results than corresponding decision trees[11] [15][15][16]. The neural network has been applied across different domains[3]. While applying a decision tree along with neural networks, we can further test how bagging and boosting are influenced by the learning algorithm, it provides a further vision of the general characteristics of these approaches. Bagging and boosting used two learning methods decision tree and Naive Bayes. The main motive is to test the performance of the decision tree.

From previous results, it's clear that the bagging ensemble method produces better results than a standard classifier. Hence, we can say that bagging is one of the safest and most effective methods to get better results for decision trees or neural networks. Boosting is unable to handle noisy data so it performs differently for different datasets. As per the author[12], boosting reduces errors when using machine learning models but in actuality, it increases the error when using the neural networks. The main reason boosting sensitivity to noisy data.

A neural network is an ensemble with other methods and tested. An ensemble of neural networks has been built where each network is trained on the same data but consists of different initial weights. Despite these different weights, which are later on combined provide good results and better performance.

The performance or accuracy of the algorithm has been improved for both neural networks and decision trees. But still, comparisons are generally done to know which performs better either bagging or boosting. Previous research shows that boosting performance depends on the type of dataset rather than the classifier type. Boosting is more sensitive to the type of data because if data is noisy then boosting will not provide accurate results. Bagging always

performs better than boosting
In this article, the investigation will be done for several classifiers required in an ensemble model for accurate prediction. Previous research shows that error starts reducing when the number of classifiers increases in an ensemble model. The next section explains the complete details of the ensemble model like bagging and boosting classifiers in the literature review. Section III represents the framework for the proposed methodology and explains results based on research. The last section includes the conclusion summarizing the findings.

2 Literature Review

Figure 1 shows the concept of the ensemble method. In this area, the neural network is combined with a random model. Each model in this diagram(M1 to MN) is trained using the training instances for that network. The output from each model is observed and at the end, it is combined with all models. The Voting will be done to find the predicted result from these models. Other researchers conclude that combining the output from different models will simply average the network's predictions[30]. Combining the output of several classifiers, if they are not ready on one decision. If all agree on one decision, there is no need to ensemble. The error rate defines the accuracy of the model. The model is accurate, if component classifiers in the example are diverse and the error rate is less than 50%[18]. According to[23], the total error of ensemble classifiers is described in two different ways as shown in Figure 2. The author focuses on applying a convolution neural network with IOT healthcare data for analyzing complex cardiovascular data effectively[20]. Convolution and recurrent neural networks are used for the prediction of various health risks. This focus is to see patterns in clinical data for making accurate predictions[24].
Average generalization error of individual classifiers: It shows how individual classifier performs individually when applied to unseen data. Every individual classifier is accurate on its own but still, they exhibit some level of error due to over-fitting or noise in the data.
Disagreement among the classifiers: When ensemble classifiers(like bagging, boosting, and stacking) that give highly accurate results also disagree with a given input. This disagreement mitigates errors of individual classifiers and also improves the performance of the classifiers.

In this study, author [28][27] also verified that the ensemble method mitigates individual bias and reduces variance, often leading to performing well on new, unseen data.
Henceforth, we will use different classifiers in the ensemble method where each classifier has a different result for prediction. Training subsets often differ for each classifier to show disagreement on predictions to improve overall accuracy.
Methods for creating an ensemble model will include those groups of classifiers(neural networks) that disagree with their predictions. Different technologies will be used during the training process like different topologies, differ-

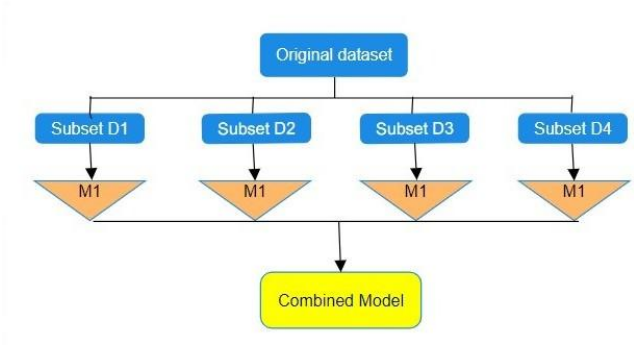


Figure 1: Ensemble Model

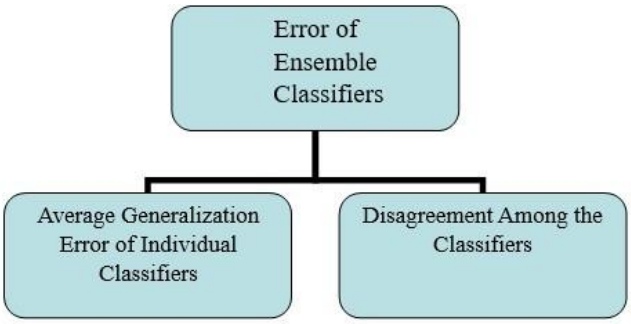


Figure 2: Measuring the Ensemble Error

ent initial weights, different parameters, and different subsets of training data [2][18]. In this article Bagging and boosting have been used for prediction.

2.1 Bagging classifiers

Bagging helps to improve accuracy by reducing variance and preventing overfitting. Without bagging, high-variance problems persist, leading to performance well on the training subset but not well on the testing subset. Randomly selected samples with replacement generate training set, N examples. The bagging resamples the training set with replacement, some instances are represented multiple times while others are left out. Training sets are resampled in bagging with replacement, where some cases are repeated again and again and some are left out as shown in Figure 3. The result from one classifier that is trained on one set generates higher test errors than the classifier using all the data. The second classifier is also trained on another set of data and gives higher test errors and so on. When we combine all the classifiers that are diverse from each other, it can produce a test-set error lower than the single classifier.

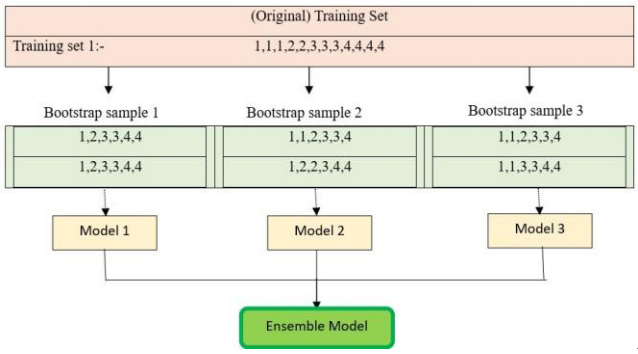


Figure 3: Bagging Method

2.2 Boosting classifiers

Boosting[12] helps reduce bias and variance to improve the model's accuracy. Mistakes or errors have been identified in previous models and the adjusting the model to correct those errors. As a result, over-fitting issues are resolved and models become stronger to handle different data samples. It improves the weakness of its predecessor model, which means learning from the past model. The training set is chosen by looking at the weaknesses and performance of the earlier classifier. To generate a more accurate model, those instances that are not correctly predicted by previous classifiers have been chosen. Hence, boosting the ability to predict new classifiers for which earlier classifiers failed to predict.

Arcing(Arcing reweighing and combining) [1] and Ada-Boosting[12] are two new and powerful forms of boosting as shown in Figure 4. For each classifier, the training set is generated to improve the performance of weak classifiers. Each iteration creates diversity in the training sets with every selection. Arcing generates the different subsets from the original data and these subsets train the model. In this way, arcing focuses on all aspects of the data and this improves the predictive performance of the ensemble model.

In the boosting process, a subset must be created with replacement from the actual data so that it covers all items. The creation of a subset helps correct errors iteratively and improves predictive accuracy. For each iteration(classifier K + 1), a new training set is created.

2.3 Trade-off between bias and variance

To measure the performance of bagging and boosting, several authors[8] [13] [21] [22] introduced several theories based n [14]trade-off between bias and variance. Bias and variance concept used in machine learning. The model is generated on training data, and tested on unseen data. Three main mechanisms could be used to test the error in the model as shown in Figure 5.

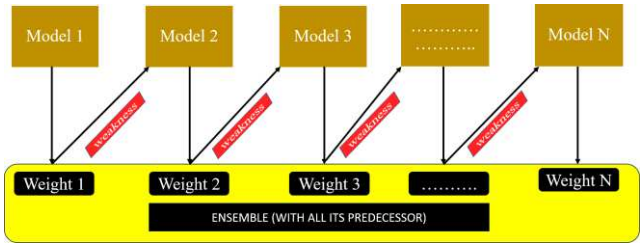


Figure 4: Boosting Classifiers

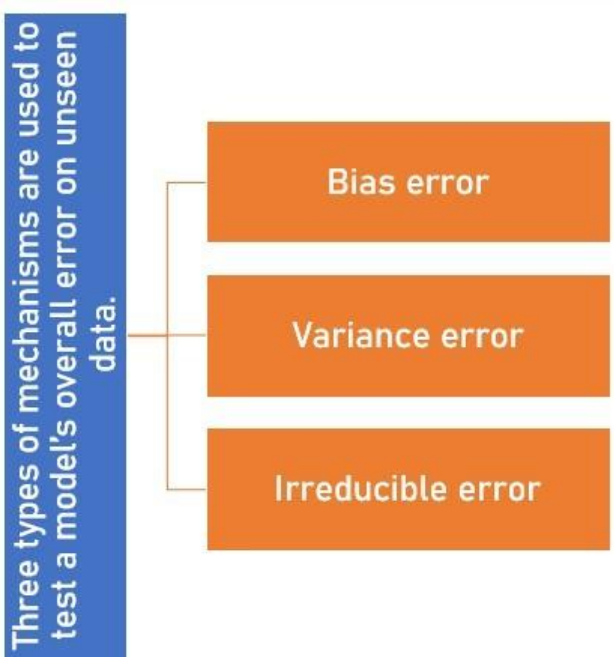


Figure 5: Mechanism to test error

Bias Error: It is helpful to predict the underlying relationships. Complex models like neural networks or decision trees must be used instead of linear prediction models. With a simple model, several significant factors will be ignored, leading to an under-fitting problem.
Variance error: High variance means the model is trained on specific data, leading to an over-fitting issue. The model didn't perform well on unseen data.
Irreducible error: Some errors cannot be reduced even if we build a better model. It is perfectly fine if the model faces minimum error.
Cardiovascular disease varies from gender to gender also. The author proved

that Myocardial infarction symptoms common in male patients(55.6%) than in female patients(32.0%)[17]. Machine learning techniques are not only helpful in the prediction of heart disease but also in another field like predicting stress levels in students by using ML techniques[25]. Deep neural networks and hybrid architectures have been combined to improve the accuracy of heart disease prediction[1].

3 Methodology

Ensemble methods like Ada-boosting, Bagging, and Arcing have been discussed in this section. Two models have been created in this study. In the first method, datasets were tested with decision trees, neural networks, and Logistic Regression. In the other method, 24 different classifiers have been used as an ensemble model for the prediction.

3.1 Data Sets

To check the model's performance, the dataset has been taken from the University of Wisconsin Machine Learning Repository and the UCI data set repository[26]. The Dataset has been chosen which consists of different characteristics:

1. (a) Related to real-world datasets.
   (b) Large size data
   (c) Mix datatype(continuous, discrete, or the mix of two)
   (d) Good outcome prediction by previous researchers.

The dataset has been shown in Table 1 that consists of details about each dataset like the name of the dataset, number of samples in the dataset, number of output classes for classification, number of continuous and discrete features, types of data, dataset size, continuous(numerical) and discrete(categorical) features, total number of inputs, outputs, hidden and epochs used for training the neural network.

Table 1: Summary of the dataset used in this paper.

Datasets	Size	Class	Features Extraction		Inputs	Outputs	Hidden	Epochs
			Cont	Discre				
Heart-Cleveland	303	2	8	5	13	1	5	40
Diabetes	768	2	9	-	8	1	5	30
Iris	159	3	4	-	4	3	5	80

3.2 Ensemble Classifier

We got the result after performing a 10-fold cross-validation test, where the same test is applied five different times. Two different models have been proposed in which one model consists of 3 different classifiers decision tree, neural network, and logistic regression(DT+NN+LR). The other model consists of a total of 24 different classifiers that have been used as an ensemble model. To test the model that consists of 24 different classifiers, data is split into 10 equal parts. Each time, 9 parts are used for training the model, and one part is used to test the model(how actual model performance). This process is repeated repeatedly so that the model must be trained using each part and each part of the data must be used as a test set exactly once. The final result is a combination of 24 different models to make a prediction. A common technique, back-propagation learning, has been used to train the neural network. Weights play an important role in back-propagation learning, where weights are going to be adjusted based on the errors made during the training phase. It's important to know parameter information like how much weights are adjusted with each update. Moderate weights are adjusted with each step. For example: Learning rate=0.15(moderate weight). But make sure it should not be zero. One more important parameter of the neural network is momentum. To make the training process works smoothly, faster and stable, the momentum value is set to 0.9. This signifies that this parameter is highly influenced by prior updates' directions. Initially, to know the impact of every input, weights are set within the range from -0.5 to 0.5 values. The number of hidden units(neurons) is decided based on the number of inputs and outputs. Different-different criteria can be chosen for hidden units and epochs as shown in Figure 6 and Figure 7.

3.3 Dataset Error Rates and Performance Metrics

The error rate shows how accurately the system performs. The model is trained using training data and to check the model's performance, the comparison is done between actual data and predicted output.

As per Table 2 and Table 3, an ensemble model has been created. The macro average precision across the two classes is 85%. These values provide an overall sense of the model's performance across both classes, without weighting by class frequency.

In Table 4, Stan(standard neural network), Sim(Simple ensemble methods), and Bag(Bagging-based neural network) shows the test set error rates for the datasets described in the table. The result concludes that each ensemble method tends to reduce the error rate for almost all of the datasets.

Table 4 shows test set error rates that show how often each model incorrectly classifies data instances. A lower value in this table indicates better

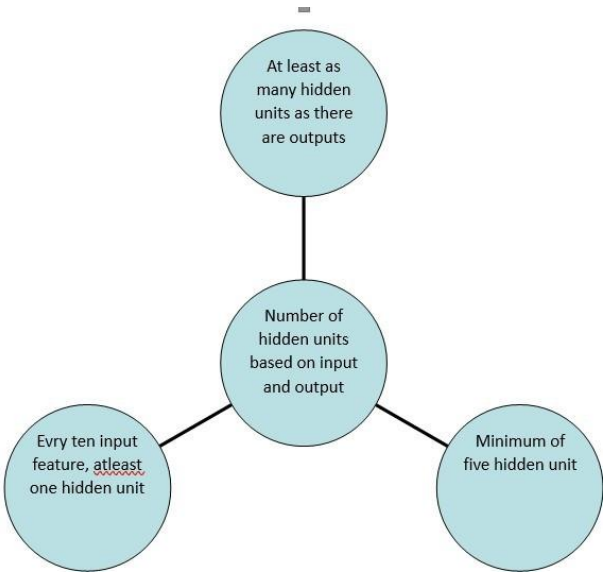


Figure 6: Criteria for choosing hidden units

Table 2: Ensemble the model using Stacking.

Datasets	Size	Class	Model	Accuracy with stack-ing	Precision with stack-ing	Recallwit stack-ing	hSupport with stack-ing
Heart-Cleveland	303	2	DT +NN+LR	86.88%	87%	87%	30.5
Diabetes	768	2	DT +NN+LR	77.97%	76%	75%	77
Iris	159	3	DT +NN+LR	100%	100%	100%	100%
Cardio vascular	70,000	2	DT +NN+LR	98%	97%	99%	40%

performance. Error rates are described for five neural network methods that explain the misclassification rates of various models. This table shows the importance of choosing the right ensemble method based on the dataset and the model used. The boosting model with arcing and Ada-boost had the lowest error, making it the most accurate for the heart Cleveland dataset. The bagging

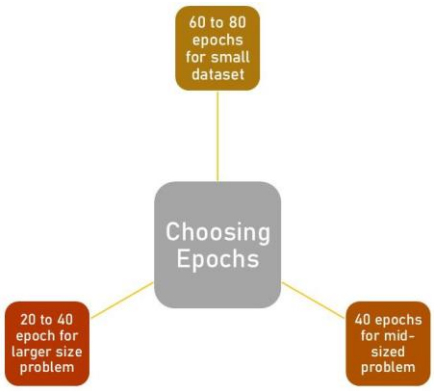


Figure 7: Criteria for choosing epochs

Table 3: Ensemble Model Using Bagging.

Datasets	Size	Class	Model	Accuracy with bagging	Avg Precision with bagging	Avg Recallwith bagging	Macro Avg Support with bagging
Heart-Cleveland	303	2	DT +NN+LR	85.24%	85%	85%	30.5
Diabetes	768	2	DT +NN+LR	73.7%	70.5%	70.5%	77
Iris	159	3	DT +NN+LR	100%	100%	100%	10
Cardio vascular	70,000	2	DT +NN+LR	70.75%	71.5%	71%	7000

method improved neural network accuracy, leading to fewer mistakes compared to other models for the diabetes dataset. The two-tailed sign test indicates that every ensemble method is significantly better than its single. The standalone neural network error rate is shown in Table 5

Figure 8 shows the reduction in error rates for two ensemble methods Adaboost and Bagging by using four different datasets.. Because of the unique characteristics of the iris dataset, this will produce low error rates that are not visually apparent. The x-axis represents different datasets like cardio, heart, diabetes, and iris and the y-axis represents error rate. The height of the bar indicates the effectiveness of the model because a lower height means the er-

Table 4: Test set error rates

Dataset	Neural Network Accuracy			Boosting Accuracy	
	Acc(Stan)	Acc(Simp	Acc(Bag)	Acc(Arc)	Acc(Ada)
Heart-Cleveland	0.15	0.15	0.15	80.3%	80.3%
Diabetes	0.29	0.26	0.31	73.3%	73.3%
Iris	100%	100%	100%	100%	100%
Cardio vascular	73.93%	74.01%	73.62%	73.66%	73.6%

Table 5: Accuracy with different classifiers.

Dataset	Neural Network Error Rate(ER			Boosting ER	
	Stan	Simp	Bag	Arc	Ada
Heart-Cleveland	85.24%	85.24	85.24%	80.3%	80.3%
Diabetes	72%	72%	68.83%	73.3%	73.3%
Iris	100%	100%	100%	100%	100%
Cardiovascular disease	73.93%	74.01%	73.62%	73.66%	73.6%

ror rate is low and a high bar means the error rate is high. When Ada-boost and bagging are both applied to the cardio dataset, Ada-boost implies lower error rates than Bagging. This indicates that Ada-boost is more suitable and effective for this dataset as compared to others. Similarly, for the heart dataset, Ada-boost performs better than bagging as it clears from the figure. But for the diabetes dataset, both ensemble models Ada-boost and Bagging perform equally because of simple data patterns or well-separated classes.

3.4 Ensemble Size

Knowing how many models we can use to reduce the test-set error is essential. Early research shows that the ten models are sufficient to reduce error. Previous studies have proved that bagging and boosting beyond ten models increase the model's performance and reduce the error[4]. Bagging and boosting are applied to neural networks and decision trees result in error reduction after adding 10 to 15 classifiers[6]. [16]proves that models perform better and better after adding up to 25 trees. The work of Ada-boost is to correct errors made by previous trees. Bagging with decision trees or neural networks, adding beyond fifteen classifiers does not improve performance. But with Ada-boost and Arcing applied to decision trees, adding more classifiers continues to improve the performance.

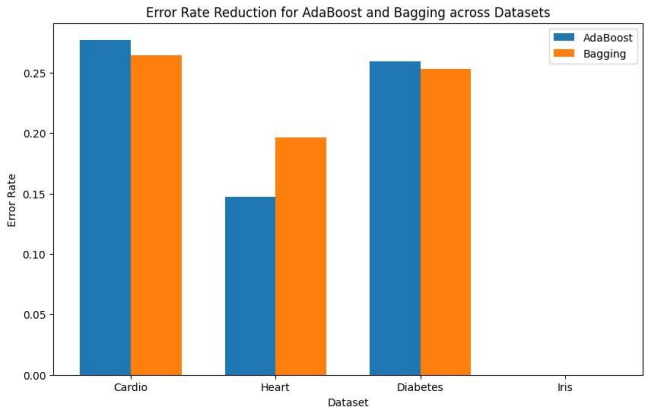


Figure 8: Error Rate Reduction for Ada Boost and Bagging across Datasets

3.5 Proposed Ensemble Model

Our proposed model MedPolyVote-24(Medical, poly means multiple, vote means voting method,24 different classifiers) for prediction of cardiovascular disease using voting method and got accuracy 74.03%, MedPolyStack-24 got accuracy 74.46%, MedPolyBagging-24 got accuracy 76.10%, MedPolyBoosting-24 got accuracy 77.37%, MedPolyRFC-24 got accuracy 78%, DT+NN+LR(Stacking) got accuracy 98%.

Table 6: Comparative analysis of ensemble models using CVD

Dataset	Model	Accuracy	Precision	Recall	fl1-Score
Cardio vascular	MedPolyVote-24	74.03%	73.88%	74.03%	73.95%
Cardio vascular	MedPolyStack-24	74.46%	74.39%	74.46%	74.42%
Cardio vascular	MedPolyBagging-24	76.10%	76.24%	76.10%	76.06%
Cardio vascular	MedPolyBoosting-24	77.37%	77.37%	77.33%	77.32%
Cardio vascular	MedPolyRFC-24	78.00%	78.04%	78%	77.99%
Cardio vascular	DT+NN+LR (Bagging)	70.75%	71.50%	71%	71%
Cardio vascular	DT+NN+LR (Stacking)	98%	97%	99%	97.90%

The ensemble model that is combination of decision tree, neural network and logistic regression performed better than other models as shown in Table 6

4 Conclusion

This study aimed at the importance of choosing the right ensemble model based on the dataset. Comparative analysis of the ensemble model is calculated by using different datasets and different numbers of classifiers. Ada-boost and Bagging both applied to different datasets for describing the error rates for five neural network methods. The boosting model with Arcing and Ada-boost has the lowest error rate, making it accurate for cardio and heart Cleveland data. But for diabetes, both performed equally. Two models have been proposed DNL in which three models are combined and Med-poly in which 24 different classifiers are combined and tested using bagging, boosting, and stacking techniques. The result shows that DNL (stacking) performed better than Med-poly. This finding suggests that these types of models perform well with healthcare data, practitioners can rely on well-constructed ensemble models. This model removes the problem of complexity and over-fitting because of a few algorithms. This model helps in the early prediction of disease where computational resources and times are constrained. The findings indicate that increasing the number of classifiers while creating an ensemble model does not always perform better. In the future, we want the focus on quality rather than quantity. Researchers can focus on optimizing the selection of classifiers rather than increasing the number of classifiers to create an ensemble model.

Funding Declaration

There is no Funding Interest from All Author

References

[1] Mana Saleh Al Reshan, Samina Amin, Muhammad Ali Zeb, Adel Sulaiman, Hani Alshahrani, and Asadullah Shaikh. A robust heart disease prediction system using hybrid deep neural networks. *IEEE Access*, 2023.

[2] Ethem Alpaydin. Multiple networks for function learning. In *IEEE International Conference on Neural Networks*, pages 9–14. IEEE, 1993.

[3] Michael A Arbib. *The handbook of brain theory and neural networks*. MIT press, 2003.

[4] Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

[5] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36:105–139, 1999.

[6] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[7] Leo Breiman. Bias, variance, and arcing classifiers. 1996.

[8] Leo Breiman. Stacked regressions. *Machine learning*, 24:49–64, 1996.

[9] Robert T Clemen. Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583, 1989.

[10] Harris Drucker and Corinna Cortes. Boosting decision trees. *Advances in neural information processing systems*, 8, 1995.

[11] Douglas H Fisher and Kathleen B McKusick. An empirical comparison of id3 and back-propagation. In *IJCAI*, volume 89, pages 788–793, 1989.

[12] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[13] Jerome H Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1:55–77, 1997.

[14] Stuart Geman, Elie Bienenstock, and Rene´ Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[15] Alan Gove. An experimental comparison of symbolic and connectionist learning algorithms. 1989.

[16] Adam J Grove and Dale Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI*, pages 692–699, 1998.

[17] Sonja Guethoff, Rebekka Kraft, Matthias Riege, Carola Grinninger, and Kara Krajewski. Females at a clear disadvantage with postoperative myocardial infarction symptoms. *Journal of Cardiovascular Development and Disease*, 11(11):371, 2024.

[18] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[19] Sherif Hashem. Optimal linear combinations of neural networks. *Neural networks*, 10(4):599–614, 1997.

[20] Aarti Karandikar, Komal Jaisinghani, Piyush K Ingole, Nilesh Shelke, Rupa A Fadnavis, and Navnath Narawade. Advanced heart disease prediction: Deep learning-enhanced convolutional neural network in the internet of medical things environment. *Journal of Electrical Systems*, 20(1s):1–10, 2024.

[21] Ron Kohavi, David H Wolpert, et al. Bias plus variance decomposition for zero-one loss functions. In *ICML*, volume 96, pages 275–283. Citeseer, 1996.

[22] Eun Bae Kong and Thomas G Dietterich. Error-correcting output coding corrects bias and variance. In *Machine learning proceedings 1995*, pages 313–321. Elsevier, 1995.

[23] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

[24] Yuxi Liu, Shaowen Qin, Antonio Jimeno Yepes, Wei Shao, Zhenhao Zhang, and Flora D Salim. Integrated convolutional and recurrent neural networks for health risk prediction using patient journey data with many missing values. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1658–1663. IEEE, 2022.

[25] Farhad Lotfi, Branka Rodic´, Aleksandra Labus, and Zorica Bogdanovic´. Smart healthcare: developing a pattern to predict the stress and anxiety among university students using machine learning technology. *Journal of Universal Computer Science*, 30(10):1316, 2024.

[26] Patrick M Murphy. Uci repository of machine learning databases. *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1994.

[27] David Opitz and Jude Shavlik. Generating accurate and diverse members of a neural-network ensemble. *Advances in neural information processing systems*, 8, 1995.

[28] David W Opitz and Jude W Shavlik. Actively searching for an effective neural network ensemble. *Connection Science*, 8(3-4):337–354, 1996.

[29] MP Perrone. Improving regression estimates: Averaging methods for variance reduction with extensions to general convex measure optimization. *PhD Thesis, Brown University*, 1993.

[30] J Ross Quinlan et al. Bagging, boosting, and c4. 5. In *Aaai/Iaai, vol. 1*, pages 725–730. Citeseer, 1996.

[31] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.