Informatics in Medicine Unlocked 26 (2021) 100655

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Early detection of coronary heart disease using ensemble techniques

Vardhan Shorewala

Dhirubhai Ambani International School, Mumbai, India

ARTICLE INFO

Keywords:
Machine learning
Heart disease
Artificial intelligence
Hybrid modelling
Coronary heart disease
Ensemble techniques

ABSTRACT

Heart disease is one the leading causes of death globally, making the early detection of it crucial. Emerging technologies such as machine learning and deep learning are now being actively used in biomedical care, healthcare, and disease prediction. The focus of this paper is on the prediction of coronary heart disease (CHD) using a risk factor approach. Predictive techniques such as K-Nearest Neighbors, Binary Logistic Classification, and Naive Bayes are evaluated on the basis of metrics such as accuracy, recall, and ROC curves. These base classifiers are compared against ensemble modelling techniques such as bagging, boosting, and, stacking. A comparitive analytical approach was used to determine how ensemble techniques can be used to improve prediction accuracy of coronary heart disease. The modelling technqiues are tested on the 'Cardiovascular Disease Dataset,' which contains 70,000 records of patient data for coronary heart disease. Bagged models are shown to have an averaged increased accuracy of 1.96% in comparison to their traditional counterparts. Boosted models had an average accuracy of 73.4% but had the highest AUC score of 0.73. The stacked model involving KNN, random forest classifier, and SVM proved to be the most effective with a final accuracy of 75.1%. In addition, the perfomance of the tested models was validated using data-analytic technqiues and K-Folds cross-validation.

1. Introduction

There have been numerous developments in the field of medical care such as fitness and health bands. Furthermore, devices such as electrocardiograms and CT scans help in the detection of coronary heart disease. However, the high cost and infeasibility of these machines are major factors that have led to the death of 17 million patients due to coronary heart disease annually [1]. Among all human diseases, the chronic disease group is considered the most dangerous as shown by a Lancet Study on Global Burden of Disease Study in 2013 [2]. The risk factors associated with the disease include excessive alcohol consumption, high blood pressure, and the sex and age of a patient. These conditions are often found in high-income countries such as the United States, where 87% of deaths were caused by chronic diseases [3]. However, a particular concern should be given to low and middle-income countries where the prevalence of chronic diseases has seen an increase in the number of cases. "In the slums of today's megacities, we are seeing non-communicable diseases caused by unhealthy diets and habits, side by side with undernutrition [4]." Conventional methods of detecting coronary heart disease include angiography which has drawbacks like high cost, various side effects, and requirements of strong technological knowledge [5].

To overcome the issues associated with conventional methods, non-invasive methods based on predictive machine learning models can be leveraged [6]. The contribution of the current work includes making an intelligent diagnostic system based on contemporary machine learning methods. In this study, 6 base models were explored: Logistic Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), Decision Tree (DT), Naïve Bayes (NB), and Neural Network (MLP). These models are then extensively compared to their ensemble counterparts on the basis of accuracy, specificity, and sensitivity to arrive at the optimum model for clinical use. The developed system and model are based on the 'Cardiovascular Heart Disease' dataset, which is publicly available on Kaggle [7]. All processing, visualization, and computation was done on Jupyter Notebooks, using python. The main contributions of this study include:

- The dataset used is relatively much larger (70,000+ values) in comparison to traditional, smaller datasets used such as the Cleveland Dataset and the Hungarian Heart Disease Dataset (200–1000 values). This helps create realistic and higher performing models
- The paper evaluates the performances of base models against their bagged counterparts. Traditional papers are limited to either one method

E-mail address: vardhanshorewala@gmail.com.

https://doi.org/10.1016/j.imu.2021.100655
Received 2 April 2021; Received in revised form 28 June 2021; Accepted 28 June 2021
Available online 11 July 2021

- The study also involves stacking and boosting. These ensemble techniques have not been explored extensively in traditional papers relating to coronary heart disease prediction
- The study includes statistical and qualitative analysis of the datasets. Pattern analysis of variables in relation to coronary heart disease prediction has not been discussed in detail in prior research.

The paper is organized as follows: a review of research related to the study is presented in Section II. A description of the dataset used in provided in Section III. Section IV presents feature analysis and examines correlations between the target variable and features. Experimental results of base and ensemble models are provided in Section V. The discussion is given in Section VI.

2. Literature review

In the growing field of data science and medical care, the need for automated diagnostic systems is increasing. Data scientists have developed several models, which have helped aid in the field of medical care. Previous studies have shown that neural networks, Naive Bayes classifiers, and associative classification are powerful methods for diagnosing coronary heart disease. This is because associative classification provides high data accuracy and data flexibility, which traditional classifiers lack [8]. In order to develop a heart disease classifier, a data mining algorithm was built for data gathering and for predictive modelling. Thousand CHD patient records were mined, and the authors used a Support Vector Machine (SVM), Artificial Neural Network (ANN), and a Decision Tree (DT) for the binary classification job. The models respectively produced accuracies of 92.1%, 91%, and 89.6%. Furthermore, K-folds validation and confusion matrices were used to evaluate the consistency, sensitivity, and specificity of the data [9]. Another data scientist used ensemble to increase data consistency and increase data accuracy. The author used bagging and boosting on Naive Bayes and Multilayer Perceptron Neural Networks. These ensemble techniques increased the accuracy by an average of 7.26% in predicting coronary heart disease. The use of SVMs in disease prediction has also proven helpful. Majid Feshki made use of Particle Swarm Optimization and Feed-Forward backpropagation neural networks to optimize features. The methods yielded an accuracy of 91.94% [10]. The K-Means Clustering was used for feature extraction from the frequent patterns that were mined using the Maximal Frequent Itemset Algorithm (MAFIA). Lastly, Muhammad, Tahir et al. conducted a comprehensive analysis of base classifiers for the prediction for coronary heart disease [11]. The Extra-Tree Classifier (ETC) proved the most effective with an accuracy of 92.09% and AUC of 97.92%. This was followed by Gradient Boosting, which had an accuracy of 91.34%. The study also highlighted the effect of feature selection algorithms such as Lasso and Relief.

A simple and reliable feature selection method was proposed to determine the heartbeat case using the weighted principal component analysis (WPCA) method. The proposed method enlarged the ECG signal's amplitude and eliminated noises, yielding an accuracy of 93.19% [12,13]. Backpropagation methods [14] help compare classification accuracies. The author delivered high accuracy output from his models. A comparative analysis [15] of accuracies on heart disease prediction used the Naive Bayes classifier, SVM, and logistic regression. The highest accuracy, 80%, was yielded by the SVM, depicting its scope in prediction. Furthermore, Nilashi et al. showcased that fuzzy SVMs with Principal Component Analysis (PCA) are able to achieve higher accuracies at predicting coronary heart disease at a lower componential time, using incremental learning [16].

Artificial Neural Networks have been employed in previous research related to heart disease prediction. Olaniyi and Oyedotun [17] proposed a three-step model based on an ANN to diagnose angina, which achieved an accuracy of 88.89%. Das et al. [18] produced an ANN ensemble-based predictive model, using a statistical analysis system. This achieved a classification accuracy of 89.01% and a specificity of 95.91%. Dutta et al. showcased that their proposed CNN architecture reached an accuracy of 77% to predict coronary heart disease and predicted negative cases with higher accuracy in comparison to traditional methods such as SVMs and Random Forests [19]. Lastly, Jabbar et al. [20] created a multilayer perceptron ANN-driven backpropagation learning algorithm and feature selection algorithm for coronary heart disease. In order to diagnose heart disease, an integrated decision support medical system based on ANN and Fuzzy Analytical Hierarchical processing was designed by the authors [21].

Clustering techniques have also been identified as helpful in diagnosing coronary heart disease [22]. Data scientists cross-compared various clustering techniques such as EM, Cobweb, K-Means, Farthest First, etc. The most effective proved to be a density-based approach to diagnosing coronary heart disease. Spectral clustering [23] has also been used in a CBIR of cardiac models [24] to help diagnose congestive heart values. The novel model yielded an accuracy of 83%.

Ensemble techniques have proved extremely powerful in predicting heart disease. A group of researchers [25] cross-compared three algorithms: c4.5, j4.8, and the bagging algorithm, and concluded that bagging was the most powerful, with an accuracy of 81.41%. This depicts the scope of ensemble techniques. Two researchers [26] combined various models and compared their respective strengths. The most powerful model was produced by combining a fuzzy Naive Bayes with a genetic algorithm. This had an accuracy of 97.14%. A group of researchers [27] helped develop a new cost function to address the limitations of the previous ensemble techniques: feature selection and low accuracy. Lastly, Baccouche et al. used an ensemble classifier with a BiLSTM or BiGRU model with a CNN model to achieve a F1 score of between 91 and 96% for prediction of heart disease [28]. The research highlighted that ensemble frameworks could overcome the problem of predicting upon an unbalanced dataset.

3. Data

The data, extracted from Kaggle, was unprocessed. The data was arranged into columns and a comma-separated values file. It contained no null values, and all variables were continuous or categorical. However, the dataset had two potent flaws. Firstly, it had a large standard deviation. The dataset was deep tailed with extreme values i.e., global anomalies. To combat this effectively and for consistency in data, the upper and lower 2% percentiles were trimmed for all continuous variables which had a high standard deviation. Furthermore, there were outliers such as where the Systolic Blood Pressure was lower than the Distal Blood Pressure. Implausible data points such as these were removed to allow for realistic data-points and predictions by the model. Lastly, the numeric variables, which were not categorical, were normalized between the range of 0 and 1 to allow for uniformity throughout the dataset. The target variable was balanced with close to an equal distribution of cases with and without coronary heart disease after preprocessing. This meant weighting was not to be applied during data analysis to balance the target variable. Post preprocessing (see Fig. 1) of continuous variables resulted in the following statistical features of the dataset [29]:

As shown in Table 1, there are 5 continuous variables in the dataset. These variables have been trimmed to remove extreme datapoints. Therefore, the range of the continuous variables are plausible and

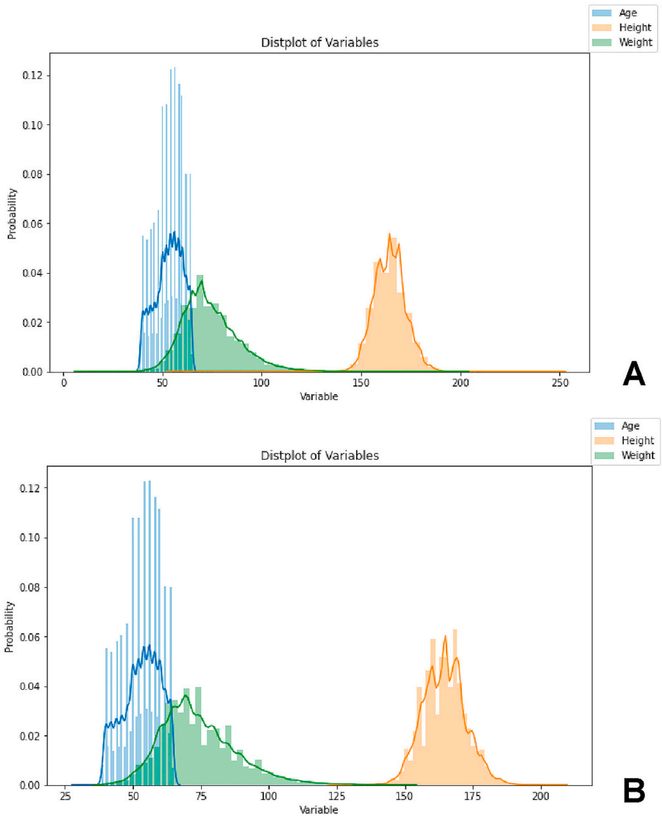


Fig. 1. The image shows the distribution of the continuous variables. **A** Distribution of variables before pre-processing. **B** Distribution of variables post pre-processing.

Table 1
Attributes of data post pre-processing.

Sr. No	Attribute Name	Range
1	Age	30 to 65
2	Height	125 to 207
3	Weight	40 to 150
4	Sex	Binary
5	Systolic Blood Pressure	70 to 240
6	Diastolic Blood Pressure	50 to 182
7	Cholesterol	1, 2 or 3
8	Glucose	1, 2 or 3
9	Smoking	Binary
10	Alcohol Intake	Binary
11	Physical Activity	Binary
12	Cardiovascular Disease	Binary

practical. On the other hand, there are 6 categorical variables, with 4 being binary. The other two categorical variables, cholesterol, and glucose, have values denoted by 1, 2, or 3. This represents the level of the attribute i.e., 3 showcases high glucose levels, while 1 shows low glucose levels. The target variable in this dataset is the cardiovascular disease i.e., a binary output.

4. Feature analysis

To examine the correlation between features and the target variable, Pearson's coefficient [30] was used to form a heatmap. To explore the relationships presented in the heatmap, similar data-points were grouped to gauge at clustering power. Age and systolic blood pressure were group with a mapping of the target variable. This helps showcase the distribution of the target variable more clearly.

Fig. 2 depicts a graph of two major continuous variables, which had high Pearson Coefficients. It maps the continuous variables against cholesterol to showcase patterns in the dataset. It made use of a random sample of 400 data-points to prevent biases in data. Fig. 2 shows that most coronary heart disease patients had higher cholesterol and systolic blood pressure. However, as Fig. 2 shows, the relationship of age with the target variable is not clear in the clustering technique employed.

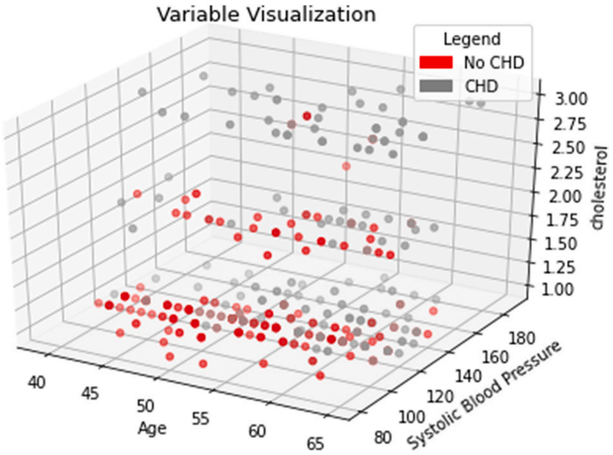


Fig. 2. 3-D graph of major variables with a mapping of the target variable.

As this task was classification based, the data-analytics explored clusters of data-pointswith a mapping of the target variable. Patterns in the data were observed using centroid-based clustering. K-Means [31] was used to calculate the position of the centroids, restricted to two dimensions for easy data visualization purposes. The centroids were plotted for the four continuous variables to identify clusters. Two graphs were produced for each set of variables: one with a mapping of the target variable and another with the clusters formed using the centroids. Fig. 3 serves as a sample, showcasing the clustering methodology for the diastolic and systolic blood pressure.

The centroid positions were calculated by:

$$X\ (data\ set)\ = \{x_1, x_2, x_3....x_c\}$$
$$V\ (cluster\ set) = \ \{v, v, v_3.....v_n\}$$
$$\sum_{i=1}^{c}\sum_{j=1}^{c_i}\left(||x_i - v_j||\right)^2$$
$$'c_i'\ is\ the\ number\ of\ date\ points\ in\ i^{th}\ cluster\ and$$
$$'c'\ is\ the\ number\ of\ cluster\ centers$$
$$'||x_i - v_j||'\ is\ the\ Euclidean\ distance\ between\ x_i\ and\ v_j$$

Fig. 3 depicts that clustering is an effective method when employed to the blood pressure variables. The diastolic and systolic blood pressure are shown as they had the highest Pearson and LASSO coefficients (see Fig. 4). As depicted in the graphs, patients with and without coronary heart disease can be sorted into clusters. The graphed data shows that clusters 5, 6, and 7 are where most of the patients having CHD lie. The other three continuous variables - age, height, and weight - were tested similarly, but they showed weaker yet significant classification. This hints at K-Means and other clustering methods working well in this classification task.

Curves of best fit were plotted for each of the 5 continuous variables. Age depicted a strong linear relationship while the rest of the variables depicted a polynomial or curved relationship. The patients diagnosed with coronary heart disease are towards the higher end of each variable. This shows that logistic classification could be powerful in classifying the data.

Lastly, to conclude data analysis and quantified feature analysis, the z-value [32] and the derived p-value were calculated for each variable, both categorical and continuous, with respect to the target variable. This would help filter variables for logistic regression but not for clustering-based models. The results showed gender to have a low Z-Value of -0.655 and a $P > |z|$ of 0.512. Therefore, gender was not

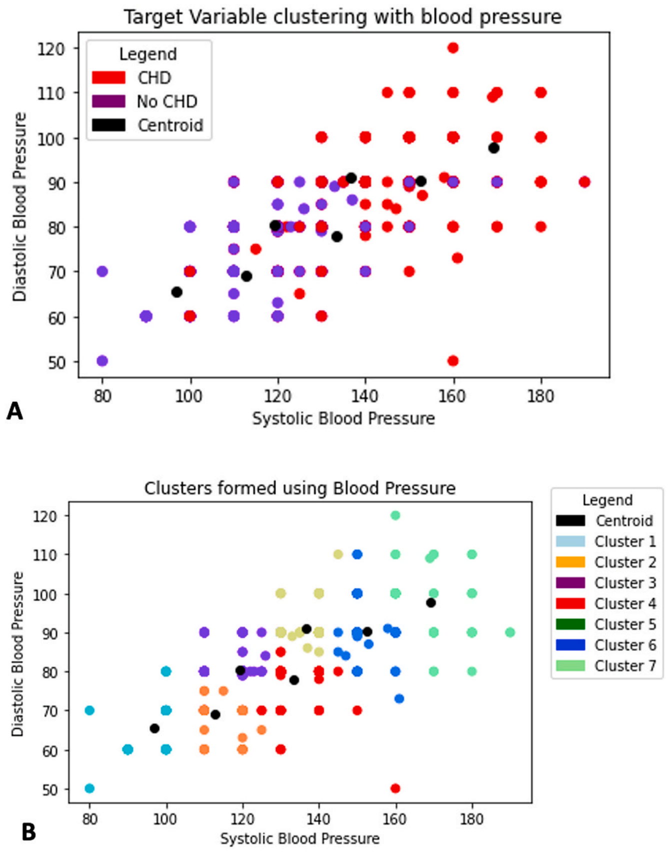


Fig. 3. Clusters formed by systolic blood pressure. **A** Clusters of datapoints formed. **B** Datapoints mapped with the target variable.

used as an input feature for logistic regression. The other variables showed high Z values, in a range of −4.21 (Height) and 60.68 (Systolic BP).

5. Experimental setup and results

After carrying out feature analysis, Least Absolute Shrinkage and Selection Operator (LASSO) was used for feature selection. Feature selection is important to remove irrelevant features that affect the classification performance of models. LASSO is based on updating the absolute value of feature coefficients. Features with lower coefficients are removed while ones with higher are retained in the dataset (see Fig. 4) .

The processed dataset was modeled using base classifiers to determine their effectiveness. The processed dataset did not include alcohol, smoking, and glucose as features as their coefficients were less than <0.01 when tested with LASSO. 75% of points were used for training the model while 25% were kept for testing evaluating performance metrics. Furthermore, to increase the randomness of the data points and to prevent overfitting, the data points were randomized in the dataset.

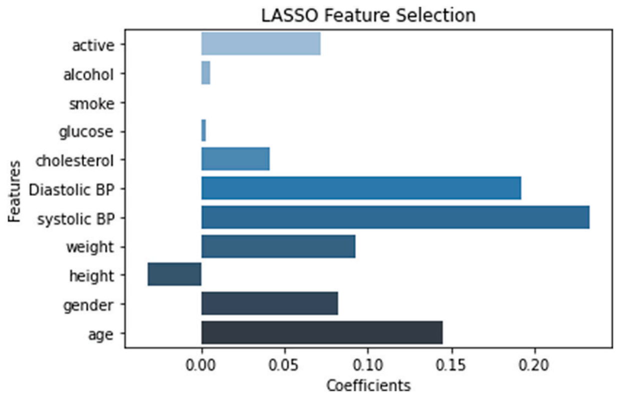


Fig. 4. Graph showing results of LASSO feature Selection.

To validate the results yielded by the models, K-Fold validation was used. In this experiment, 10 K-Folds was used. The results of each performance metric was averaged and returned for the respective model. Lastly, using nested loops, ranging over large values, the major hyper-parameters such as the verbosity, iterations, and leave nodes were optimized to give the optimum results. Grid search was then used to get the optimum combination of hyper-parameters for each model tested. The algorithm was created using python and sci-kit learn was used for the modelling of the dataset.

5.1. *Traditional classifiers*

Table 2 shows that base classifiers had a similar level of accuracy, with an average of 72.3%. The neural network had the highest accuracy, 73.93%, coupled with the greatest AUC (see Fig. 5).

The decision tree (D_1) proved to be the most successful base classifier, with an accuracy of 73%. Although having received a high AUC score, it lacked a high recall score. This means the actual class of having coronary heart disease was predicted incorrectly. Therefore, a modified version of the decision tree (D_2) was proposed, giving an equal balance between the recall and precision but with significantly lower accuracy of 71.4%. However, it had a more balanced recall and precision score of 72.6% and 71.7% respectively. This modified decision tree had a larger number of leaf nodes per branch in comparison to the prior one.

Table 2
Performance metrics of traditional classifiers.

Model	Accuracy	Recall	Precision	10-K Folds STD	F1 Score	AUC
Logistic Classification	71.4	67.8	72.6	0.58	70.1	0.72
K-Nearest Neighbors	71.6	66.6	73.3	0.43	69.8	0.71
Decision Tree (J48)	73.0	63.3	77.8	0.56	69.8	0.72
SVC	72.2	62.9	76.7	0.49	69.1	0.72
Gaussian Naive Bayes	71.4	61.2	76.3	0.84	67.9	0.70
Neural Network (MLP)	73.9	69.1	75.2	0.62	72.0	0.73

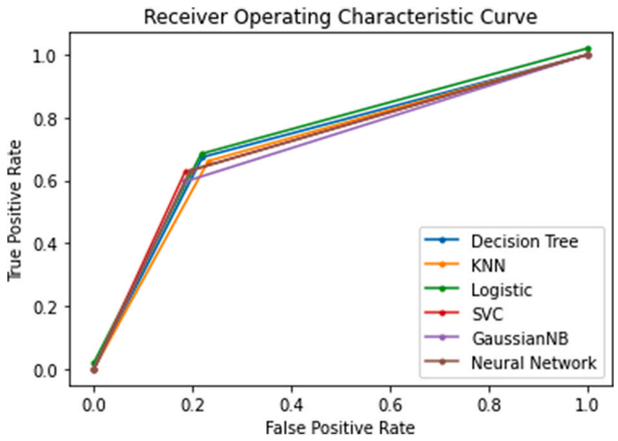


Fig. 5. Receiving operator curves for base classifiers.

5.2. Dense neural network

A dense neural network is a supervised learning algorithm, which modifies weights and biases during training to achieve at the optimum combinations. The MLP had a fully connected neurone architecture with a low learning rate to allow the model to arrive at the global optimum set of weights and biases.

Proposed architecture:
Dense Input layer, 128 neurons (input dim = 11)
Batch Normalization, Batch Dropout (0.6)
Dense Hidden layer, 256 neuron, activation = 'sigmoid'
Batch Normalization, Batch Dropout (0.3)
Dense Hidden layer, 256 neuron, activation = 'SoftMax'
Batch Normalization, Batch Dropout (0.15)
Dense Hidden layer, 256 neuron, activation = 'sigmoid'
Batch Normalization
Dense Output layer, activation = 'relu' (output dim = 2)
***Optimizer**: Adam, Adaptive, Dynamic Learning Rate of 0.01*
***Loss function**: Categorical cross entropy*

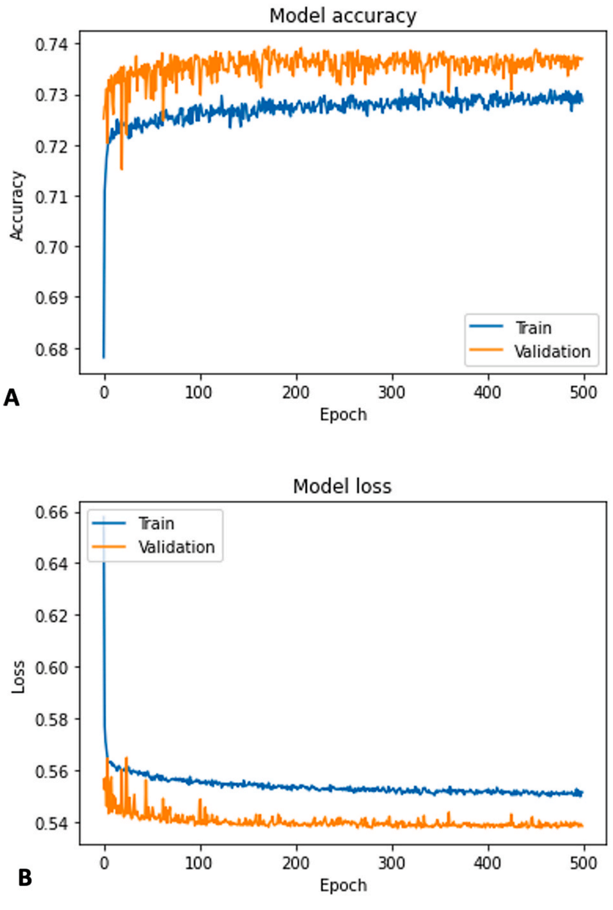


Fig. 6. Evaluation of neural network. **An** Accuracy of network with epochs. **B** Loss of model with varying epochs.

The neural network was trained for over five-hundred epochs, as shown in Fig. 6. The multi-layer perceptron based neural network showed the highest testing accuracy of 73.9% and a F1 score of 72.0%. The score was achieved for 172 epochs, where the loss function reached a global minima of 0.5380.

5.3. Ensemble techniques

Ensemble techniques aim to reduce variance across a singular model by combining various heterozygous or homozygous models. In boosting, a strong classifier is built by combining several different weak classifiers with an iterative process. However, bagging is a homogenous technique where base classifiers are fitted on various subsets of the data to help aggregate their performance. Furthermore, random forests will be explored, a bagging algorithm as it fits various subsets of the data over multiple decision trees. The number of estimators was varied across each ensemble model to gauge the optimum number.

5.3.1. Boosting

Boosting is a homogenous technique, where the base classifier is trained upon subsets of data to help produce several models of moderate performance. The wrongly classified data points are then sorted into subsets and fitted to the next model. Therefore, by combining various weak learners, using a cost function, the variance of the model is decreased. The base estimator used was the default i.e., the tree algorithm, CART. The major hyperparameters such as estimators and the number of times the model is boosted was varied through an iterative process to reach the optimum number. The best combination of hyperparameters was then chosen through grid searching.

Fig. 7 depicts that the Gradient boosting algorithm was the most effective from the three boosted models evaluated while the least effective method proved to be the ADA Booster. Furthermore, the number of estimators was directly proportionate to the accuracy until 150 estimators. For estimators greater than 150, the accuracy gradually decreased for the models except for the XGB Booster (see Fig. 7).

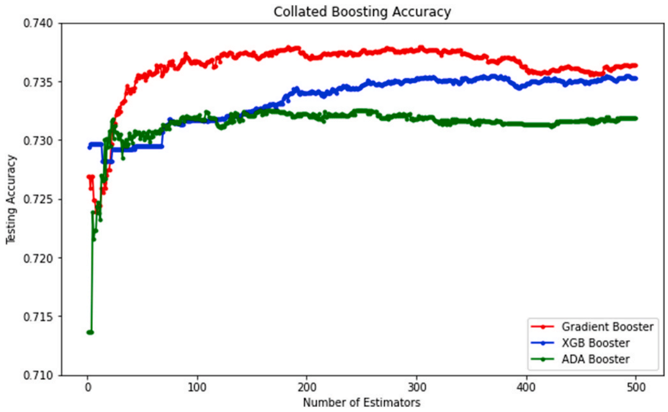


Fig. 7. Boosted models' accuracy with varying number of estimators.

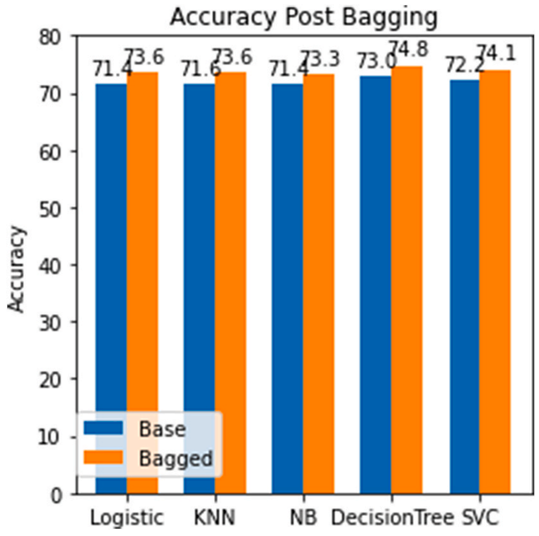


Fig. 8. Accuracy comparison of bagged and base models.

5.3.2. Bagging

Bagging involves generating multiple versions of a predictor and using them to get an aggregated predictor. Using plurality voting, the aggregation averages the different versions of the predictor. This increases the performance of a weak classifier by using homogenous models in parallel and then averaging the outcome using a function. Each model was bagged, and the cross compared to the original base model.

The results show that the ensemble technique of bagging was effective and increased the accuracy of each model by at least +1.8% (see Fig. 8). Furthermore, the recall and precision scores were also increased by a positive factor. This means that the number of false positives and negatives are reduced, and the overall performance of the model is bettered.

Following this, the random forest model was explored. It fits data on multiple decision trees and averages the bias throughout the models. This prevents overfitting of the data on a singular decision tree, while also decreasing variance within the data. However, the issue with the random forests is the complexity involved in the model, which increases the computational time in comparison to a singular decision tree. The random forests were pruned and the number of base estimators, the number of trees, was varied to arrive at the optimum number. The optimum number of trees was 147, yielding an accuracy of 74.42%.

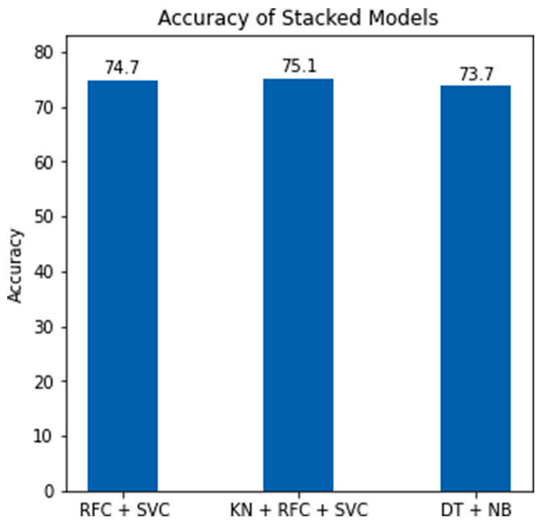


Fig. 9. Accuracy of best stacking models.

Table 3
Statistical evaluation of bagged and boosted models.

Model	Accuracy	Recall	Precision	F1 Score	AUC
Bagged Logistic Classification	73.6	68.2	75.2	71.5	0.71
Bagged K-Nearest Neighbors	73.6	66.6	73.5	69.9	0.72
Bagged Decision Tree (J48)	74.8	67.4	76.2	71.5	0.73
Bagged SVC	74.1	63.4	76.9	69.5	0.72
Bagged Gaussian Naive Bayes	73.3	61.1	76.2	67.8	0.70
Bagged Random Forest	74.4	67.3	76.6	71.2	0.73
XGB Boost	73.6	73.56	75.95	71.7	0.74
Gradient Boosting	73.2	73.79	75.35	72.4	0.73
AdaBoost	73.5	73.26	76.92	70.7	0.73

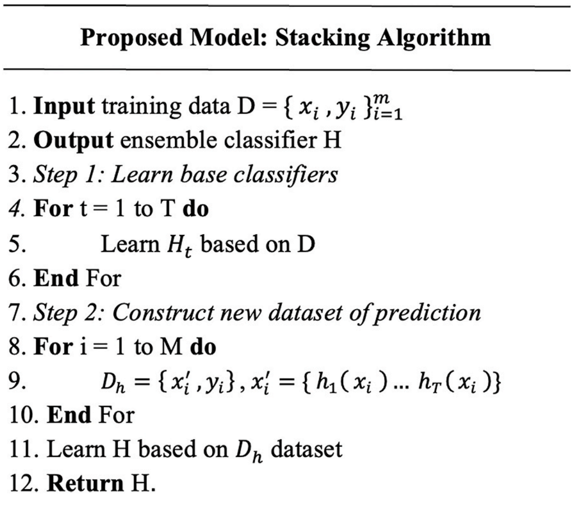
5.3.3. Stacking

The last ensemble technique explored was stacking, which is a powerful modelling method, combining heterogeneous weak learners. Multiple layers are created, where each model passes their results to the layer on top. The topmost layer makes the final decision while the bottom-most layer receives the inputs from the original dataset. The meta classifier used for combining the different classifiers is the majority vote. The topmost layer, or the final model, is the binary logistic classifier as that proved to be the most effective. The set of base classifiers available for stacking is shown below.

Models = {Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbors, SVC}

To arrive at the optimum stacked modelling, each subset of models was created using backtracking and then fitted to the stacked model.

The stacking of KNN, random forest classifier, and support vector machine with logistic regression as the meta-classifier produced the highest accuracy of 75.1% (see Fig. 9).



Proposed Model: Stacking Algorithm

```
Input training data D = { $x_i$ , $y_i$ }$_{i=1}^{m}$
Output ensemble classifier H
Step 1: Learn base classifiers
For t = 1 to T do
    Learn $H_t$ based on D
End For
Step 2: Construct new dataset of prediction
For i = 1 to M do
    $D_h = \{x_i', y_i\}, x_i' = \{h_1(x_i) \dots h_T(x_i)\}$
End For
Learn H based on $D_h$ dataset
Return H.
```

6. Discussion

Current research mainly focuses upon traditional classifiers. The only ensemble techniques explored for diagnosing coronary heart disease are boosting and random-forest classifiers. This study helps showcase how stacking and bagging are effective and more reliable methods than the ones currently being tested. On a study on the Cleveland dataset, the random-forest classifier and decision tree were found to be the most effective [33]. As shown in Table 3, the random forest classifier and decision tree proved to be the most effective in our study as well with accuracies of 74.4% and 74.8% respectively. Previous studies have also showcased how bagged models are more effective to their traditional counterparts [34], which is clearly shown in the results yielded by this study as well. Several studies that used ensemble techniques for prediction of coronary heart disease support the fact that boosted models and random forests outperform base classifiers significantly [35–37]. The boosted models employed outperformed the base classifiers for each metric evaluated in this study as well. This included AdaBoost, Gradient Boosting, and XGBoost.

This study explores stacking as an alternative to traditional methods for the prediction of coronary heart disease in patients. As shown in the results, stacking of models proved to have the highest accuracy as compared to base-classifiers and other ensemble techniques. This alternative has not been explored extensively in previous literature related to predicting coronary heart disease. Although previous studies provide models with greater accuracies, their datasets are significantly smaller than the one explored in this study. This renders most previous models impractical with real data. However, the proposed model deals with a large dataset, making the proposed model more practical, efficient, and robust.

7. Conclusion

This research analyzes the effectiveness of machine learning in predicting coronary heart disease. Firstly, the data analytics revealed patterns in the data and important features for the binary logistic classification. The statistical approach in addition to the k-nearest neighbors played a vital part and allowed for effective feature selection from the dataset. The models explored, however, had a capped accuracy at 75%. The base models analyzed had an average accuracy of 71.92% while the neural network approached 73.97% accuracy. The ensemble techniques of bagging, boosting, and stacking proved effective in raising the accuracy of the base models. The average accuracy change shown by bagging was +1.9%, raising the bagged models' average accuracy to 73.82%. While boosting, the Gradient Boosting approach proved to be the most powerful, yielding an accuracy of 73.89% when optimized. The K-Folds Cross-Validation depicted the consistency in the results produced with models: accuracies had low standard deviations ranging from 0.3% to 0.6%. Stacking, involving heterozygous models, proved to be the most effective ensemble method, producing an accuracy of 75.1%. This involved stacking KNN, random forest classifier, and support vector machine with logistic regression as the meta-classifier.

However, the other statistical methods depicted flaws in the models. The precision of each model was high, with an average of 76.1%. However, the recall score was considerably lower, which decreased the area under the Receiving Operator Curves as well. An average of 66.8% was achieved for the recall score across all models. In the future, we aim to validate the proposed model on lab test data to gauge the practicality of the predictions. Additionally, other ensemble techniquessuch as ensemble neural networks will be explored. Currently, the study was limited to ensemble techniques such as boosting, bagging, and stacking.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] The atlas of heart disease and stroke. 2010, December 09. Retrieved from, http://www.who.int/cardiovascular_diseases/resources/atlas/en/.

[2] Huse O, Gearon E, Peters A. Obesity in Australia. 2017, October 31. Retrieved from, https://pubmed.ncbi.nlm.nih.gov/29097148/.

[3] Schmidt H. Chronic disease prevention and health promotion. 2016, April 13. Retrieved from, https://www.ncbi.nlm.nih.gov/books/NBK435779/.

[4] Retrieved from, http://www9.who.int/whr/2002/message_from_the_director_general/en/; 2002.

[5] Gonsalves AH, Thabtah F, Mohammad RM, Singh G. Prediction of coronary heart disease using machine learning. Proceedings of the 2019 3rd international conference on deep learning. Technologies - ICDLT 2019. https://doi.org/10.1145/3342999.3343015.

[6] Patil SB, Kumaraswamy Y. Intelligent and effective heart attack prediction system using data mining and artificial neural network. Eur J Sci Res 2009;31:642–56.

[7] Svetlana ulianova. Cardiovascular disease dataset. Retrieved from, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset; 2019, january 01.

[8] Latha CB, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. July 02, https://www.sciencedirect.com/science/article/pii/S235291481830217X; 2019. Retrieved from.

[9] Lv J, Zhang X, Han X, Fu Y. A novel adaptively dynamic tuning of the contention window (CW)for distributed coordination function in IEEE 802.11 ad hoc networks. 2007 international Conference on convergence information technology (ICCIT 2007). 2007. https://doi.org/10.1109/iccit.2007.146.

[10] Khazaee A. Heartbeat classification using Particle Swarm optimization. Int J Intell Syst Appl 2013;5(6):25–33. https://doi.org/10.5815/ijisa.2013.06.03.

[11] Muhammad Yar, Tahir Muhammad, Hayat Maqsood, Chong Kil To. Early and accurate detection and diagnosis of heart disease using intelligent computational model. Sci Rep 2020;10(1):1–17.

[12] Yeh Y, Chen C, Chiou CW, Chu T. A reliable feature selection algorithm for determining heartbeat case using weighted principal component analysis," 2016 International Conference on System Science and Engineering. Puli: ICSSE); 2016. p. 1–4. https://doi.org/10.1109/ICSSE.2016.7551594.

[13] Dubey VK, Saxena AK. Hybrid classification model of correlation-based feature selection and support vector machine," 2016 IEEE International Conference on Current Trends in Advanced Computing. Bangalore: ICCTAC); 2016. p. 1–6. https://doi.org/10.1109/ICCTAC.2016.7567338.

[14] Al-Milli, Nabeel R. Back propagation neural network for prediction of heart disease. J Theor Appl Inf Technol 2013;56:131–5.

[15] Srinivasaraghavan Anuradha, Joseph Vincy. Comparative analysis of accuracy on heart disease prediction using classification methods. International Journal of Applied Information Systems 2016;11:22–5. https://doi.org/10.5120/ijais2016451578.

[16] Nilashi Mehrbakhsh, Ahmadi Hossein, Azizah Abdul Manaf, Rashid Tarik A, Samad Sarminah, Shahmoradi Leila, Aljojo Nahla, Akbari Elnaz. Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates. Int J Fuzzy Syst 2020:1–13.

[17] Olaniyi EO, Oyedotun OK. Heart diseases diagnosis using neural networks arbitration. Int J Intell Syst Appl 2015;7(12):75–82.

[18] Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 2009;36(4):7675–80.

[19] Dutta Aniruddha, Batabyal Tamal, Basu Meheli, Scott T. Acton. "An efficient convolutional neural network for coronary heart disease prediction. Expert Syst Appl 2020;159:113408.

[20] Jabbar MA, Deekshatulu BL, Chandra P. Classification of heart disease using artificial neural network and feature subset selection. Global Journal of Computer Science and Technology Neural & Artificial Intelligence 2013;13(11).

[21] Samuel OW, Asogbon GM, Sangaiah AK, Fang P, Li G. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Syst Appl 2017;68:163–72.

[22] Pandey AK, et al. Datamining clustering techniques in the prediction of heart disease using attribute selection method. Heart Dis 2013;14.

[23] Andrew Y. Ng, Michael I. Jordan, Yair Weiss. "On spectral clustering: analysis and an algorithm." Adv Neural Inf Process Syst.

[24] Bergamasco LCC, Oliveira RAP, Wechsler H, Dajuda C, Delamaro M, Nunes FLS. Content-based image retrieval of 3D cardiac models to aid the diagnosis of congestive heart failure by using spectral clustering," 2015 IEEE 28th international symposium on computer-based medical systems. Sao Carlos; 2015. p. 183–6. https://doi.org/10.1109/CBMS.2015.71.

[25] Shouman M, et al. Using decision tree for diagnosing heart disease patients. Proceedings of the ninth australasian data mining conference-volume 121. Australian Computer Society, Inc.; 2011. p. 23–30.

[26] Singh N, et al. "Heart disease prediction system using hybrid technique of data" mining algorithms. International Journal of Advance Research, Ideas and Innovations in Technology 2018;4(2):982–7.

[27] Nourmohammadi-Khiarak J, Feizi-Derakhshi M, Behrouzi K, et al. New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. Health Technol 2020;10:667–78. https://doi.org/10.1007/s12553-019-00396-3.

[28] Baccouche Asma, Garcia-Zapirain Begonya, Cristian Castillo Olea, Adel Elmaghraby. Ensemble deep learning models for heart disease classification: a case study from Mexico. Information 2020;11(4):207.

[29] Čisar Petar, Čisar Sanja. Skewness and kurtosis in function of selection of network traffic distribution. Acta Polytechnica Hungarica 2010;7.

[30] Sedgwick Philip. Pearson's correlation coefficient. BMJ 2012;345:e4483. https://doi.org/10.1136/bmj.e4483. e4483.

[31] Jain Mrs, Gupta Prof. A review and analysis of centroid estimation in k-means algorithm. IJARCCE 2018;7:42–6. https://doi.org/10.17148/IJARCCE.2018.789.

[32] Behboodian J, Asgharzadeh Akbar. On the distribution of Z-scores. Iran J Sci Technol 2008;32:71–8. Transaction A.

[33] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 2019;7:81542–54. https://doi.org/10.1109/ACCESS.2019.2923707.

[34] Vu T, Braga-Neto U. Is bagging effective in the classification of small-sample genomic and proteomic data? J Bioinform Sys Biology 2009:158368. https://doi.org/10.1155/2009/158368. 2009.

[35] Tama Bayu Adhi, Lee Seungchul, Im Sun. Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. BioMed Res Int 2020:9816142. https://doi.org/10.1155/2020/9816142.

[36] Krittanawong C, Virk HUH, Bangalore S, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 2020;10:16057. https://doi.org/10.1038/s41598-020-72685-1.

[37] Sultan Bin Habib A-Z, Tasnim T, Billah MM. A study on coronary disease prediction using boosting-based ensemble machine learning approaches," 2019 2nd international conference on innovation in engineering and technology (ICIET). 2019. p. 1–6. https://doi.org/10.1109/ICIET48527.2019.9290600. Dhaka, Bangladesh.

Update

Informatics in Medicine Unlocked

Volume , Issue , , Page

DOI: https://doi.org/10.1016/j.imu.2024.101598

ARTICLE IN PRESS

Informatics in Medicine Unlocked xxx (xxxx) xxx

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Corrigendum to “Early detection of coronary heart disease using ensemble techniques” [Inform Med Unlocked 26 (2021) 100655]

Vardhan Shorewala [a,*], Shivam Shorewala [b]

[a] Dhirubhai Ambani International School, India
[b] University of California, Berkeley, United States

The authors would like to correct the author list of the original article to include Shivam Shorewala, who made substantial contributions to the conception of the work, wrote stacking algorithms essential for the paper’s arguments, and contributed to the critical revision of the paper and approved the final version submitted. The authors would like to apologise for this oversight and for any inconvenience caused.

DOI of original article: https://doi.org/10.1016/j.imu.2021.100655.

* Corresponding author.
E-mail addresses: vardhanshorewala@gmail.com (V. Shorewala), shivamshorewala@berkeley.edu (S. Shorewala).

https://doi.org/10.1016/j.imu.2024.101598

Available online 30 October 2024
2352-9148/© 2024 Published by Elsevier Ltd.