Journal of Theoretical and Applied Information Technology
30th September 2024. Vol.102. No. 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AMALGAMATE APPROACHES CAN AID IN THE EARLY DETECTION OF CORONARY HEART DISEASE

VENKATESWARA RAO CHEEKATI [1],Dr.KRISHNA MOHAN KAJA[2] ,RAJA RAO PBV[3]

GARIMIDI SUBBARAO[4],Dr.AYYAPPA CHAKRAVARTHI M[5]

[1]Asst.Professor, Koneru Lakshmaiah Education Foundation, Department of CSE,Vaddeswaram,AP,India
[2]Assoc.Professor,RV Institute of Technology,Department of ECE,Guntur,AP,India
[3]Assoc.Professor,Shri Vishnu Engineering College for Women(A),Dept.of CSE,Bhimavaram, AP,India.
[4]Asst.Professor,PACE Institute of Technology &Science,Dept.of CSE(AI&DS),Ongole,AP,INDIA.
[5]Assoc.Professor,KKR & KSR Institute of Technology and Sciences,Dept.of CSE-DS,Guntur, AP,India.
E-mail:[1]chvraograce@gmail.com,[2] krishnamohan506@gmail.com,[3]rajaraopbv@gmail.com
[4]subbug81@gmail.com, [5]ayyappam.csd@kitsguntur.ac.in

ID 55308 Submission	Editorial Screening	Conditional Acceptance	Final Revision Acceptance
10-08-24	16-08-2024	11-09-2024	23-09-2024

ABSTRACT

Due to the fact that heart disease is the leading cause of death worldwide, it is critical to recognize it early. Artificial intelligence (AI) is a relatively new technology that is being heavily applied in a variety of fields, including biomedical care and disease prediction. Deep learning and machine learning are two examples of relatively new technologies that are being heavily applied in the fields of biomedicine, healthcare, and the early detection of disease. The goal of this study is to see if human coronary heart disease risk factors can be predicted using risk variables (CHD). In order to evaluate the effectiveness of prediction techniques like K-Nearest Neighbors, Binary Logistic Classification, and Naive Bayes, it is required to measure the accuracy and recall of each prediction method (BLC). Bundling and boosting are examples of ensemble modelling techniques are comparable to these methods of predicting the future. For the purpose of determining whether or not ensemble techniques can improve the accuracy of coronary heart disease prediction, a comparative analytical method was adopted. These patient data records for coronary heart disease total approximately 70,000 records and serve as a testing ground for the modeling methodologies that are currently being researched and developed. There is a 1.96 percent increase in accuracy between bagged models and their conventional equivalents. The improved models outperformed all other models by a wide margin, with an average AUC of 0.73. A combined accuracy of 75.1 percent was achieved by using the SVM, KNN, and random forest classifiers, which were regarded to be the most accurate. Utilizing data analysis and K-Fold cross-validation, the performance of the tested models was assessed

Keywords: *Heart Disease, Hybrid Modeling, Artificial Intelligence, Ensemble Method, Machine Learning And Coronary Heart Disease*

1. INTRODUCTION

Fitness and health monitors are only a few instances of noteworthy recent medical sector advances. Electrocardiograms and CT scans, as well as other diagnostic methods, can also be used to diagnose coronary heart disease. Numerous factors contribute to the 17 million fatalities each year

caused by coronary heart disease, including exorbitant costs and difficulty to operate the machines. [2] Chronic diseases are the most lethal of all human ailments, according to the The Lancet Global Burden of Disease Study, which was published in the journal The Lancet in 2013. This illness can occur as a result of excessive alcohol use, high blood pressure, or simply being a certain gender or age.. Predisposition to these diseases is seen in high-income countries, such as the United States, and they account for 87 percent of all deaths worldwide [3,4]. Low- and middle-income countries are experiencing an increase in the prevalence and incidence of chronic diseases, making it vital to keep track of their prevalence and incidence in these regions. For example, we are witnessing an increase in no communicable diseases in today's megacities, which is a result of poor dietary and lifestyle choices, as well as malnutrition. A significant disadvantage of existing methods of diagnosing coronary heart disease is that they are expensive, have several

harmful effects, and necessitate a high level of technical knowledge. Traditional approaches are being phased out in favour of more innovative ones, rather than vice versa.

The use of intrusive approaches based on machine learning algorithms that predict outcomes is possible in the treatment of some illnesses. Based on cutting-edge machine learning techniques, this research led in the development of an intelligent diagnostic system for use in hospitals. When it came to the study's six essential models, researchers looked at the following algorithms: LR, SVM, KNN, Decision Tree, Nave Bayes, and Neural Networks, to name a few (MLP). These models are subjected to thorough comparisons with their ensemble equivalents in order to decide which model is the most appropriate for clinical use. The dataset used in the development of the system and model was Kaggle's 'Cardiovascular Heart Disease.' [7] The Jupyter Notebooks and Python programming environments were used to process, visualise, and compute the entire project. The following are the most noteworthy findings of the study, in no particular order.

Comparing the dataset used with smaller datasets such as the Cleve- land Dataset or the Hungarian Heart Disease Dataset (both of which contain between 200 and 1000 variables), the dataset used is significantly larger (70,000). In the end, it is possible that more realistic and useful models will emerge. It compares base models to their bundled counterparts in this study. The method described here is the only one that can be used in traditional publishing because it is the only one that is available.

The study also includes investigations into the techniques of stacking and boosting. The use of ensemble techniques has been limited in conventional coronary heart disease research due to the fact that they are a relatively new technique.

The datasets in this study are subjected to quantitative and qualitative analysis. Prior to this work, there was no comprehensive investigation into how to forecast coronary heart disease through the use of pattern analysis of indicator data.

The following are the sections that make up the overall structure of the paper: introduction, body, and conclusion Using a literature review format, Section II of the study gives a comprehensive overview of pertinent material. Section III of the paper contains a full overview of the datasets used in the investigation. Section IV attribute analysis that takes a feature-based approach looks for connections between the target variable and its attributes. Section V presents simulations using both base and ensemble models, and Section VI presents the outcomes of these simulations. The document's conclusion can be found in Section VI.

.

2. LITERATURE REVIEW

The domains of data science and healthcare, among others, are increasingly in need of automated diagnostic tools. Data scientists have made a significant contribution to the area of medicine through their models. Associative categorization and associative neural networks have been shown to be efficient in detecting coronary heart disease in a previous study. An example of this is shown by the fact that associative categorization is more accurate and adaptive than conventional classification: In order to create a system for classifying cardiac diseases, data mining techniques were used to collect data and build prediction models. Artificial neural networks, support vector machines, and decision trees were utilised to categorise thousands of CHD patient records. The models predicted the weather with 92.1 percent, 91 percent, and 89.5 percent accuracy, respectively. An evaluation of the accuracy of the data was done using K-folds validation and confusion matrixes [9]. Ensemble was used by another data scientist to improve the consistency and accuracy of data they had previously collected. According to the author, Naive Bayes and Multilayer Perceptron Neural Networks' performance was improved using bagging and boosting. An average of 7.26 percent improvement in the ability to accurately predict coronary heart disease was achieved by using these ensemble techniques. Predicting disease using statistically significant variance (SVM) has been shown to be beneficial. Particle swarm optimization and feed-forward back propagation neural networks were employed by Majid Feshki to improve the features of his model, which he reported in Nature Communications. The strategies were judged to be 91.94 percent successful based on the findings. Maximal Frequent Item Set Method and the K-Means Clustering approach were used combined to extract features from often occurring patterns that were observed (MAFIA). Base classifiers for predicting coronary heart disease were found to be successful by Muhammad, Tahir, and their colleagues [11]. ETC was found to be most accurate classifier, with an accuracy of 92% and an AUC of 97%, according to the researchers, who used the study's data. When it came time to test the next strategy, gradient-based boosting was shown to be 91.34% effective. Other methods of selecting features like Lasso and Relief were also investigated.

For the heartbeat scenario, we developed methods based on weighted principal component analysis (PPA) (WPCA). ECG signal amplitude was enhanced, while noise was reduced, leading to a precision of 93.19% [12,13]. In order to compare different algorithms, back propagation methods [14] are useful. With the help of his models, the author was able to produce high-quality results, which he then presented to the audience in an engaging manner. The Naive Bayes classifier and the SVM were both used to predict heart disease in a comparison study [15]. When it came to predicting future events, the SVM had an 80 percent success rate. Nilashi et al. added to the body of evidence supporting the usefulness of PCA-based fuzzy SVMs by showing that incremental learning can reliably anticipate coronary heart disease with reduced componential times [16]. Artificial neural networks previously allowed for this to be accomplished (ANNs)

It is important to examine the overall prognosis of cardiovascular disease while making a decision. The accuracy of Olaniyi and Oyedotun's three-step technique for detecting angina climbed to 88 percent when combined with earlier tests, according to their findings. Das and colleagues [18] used a statistical analytic method to build an ANN ensemble-based predictive model, which was then tested. With an accuracy rate of 89% and a specificity rate of 9591%, both above average, this technique yielded excellent results. Accuracy in the prediction of coronary heart disease has been proven to be 77%, while also being more accurate than current approaches like SVMs and Random forests, among other results obtained by the proposed CNN architecture Artificial neural network (ANN)-driven backpropagation learning was used to treat coronary heart disease in a study by Jabbar et al. [20]. The authors used ANN and Fuzzy Analytical Hierarchical processing to construct a medical decision support system for the detection of cardiac disease [21].

Many studies have shown that clustering algorithms can be useful in detecting coronary heart disease [22], as previously indicated. In order to find the best clustering method, we compared EM, Cobweb, K-Means, Farthest First, and other similar algorithms. For detecting coronary heart disease in the general population, the density-based technique is the most effective method, according to research. An efficient method for detecting congestive heart failure has been found in studies using spectral clustering in CBIR cardiac models [24]. Over the prior model, the new one had an accuracy rate of 83%, which was an enormous improvement.

Ensemble approaches have been proven to be incredibly beneficial when it comes to forecasting heart illness. C4.5 algorithm, J4.8 algorithm and bagging algorithms were all tested in this experiment. When compared to the other approaches, bagging had the highest accuracy rate of 81.41 percent, according to the study's findings. Improvisation and other abilities often employed in ensembles will be demonstrated using this way. An extensive study conducted by two experts [26] examined a wide range of models to determine the advantages and disadvantages of each. The most powerful model can be developed by combining a fuzzy Naive Bayes model with a genetic algorithm. The overall success rate of 97.14 percent was attained as a result of these efforts. A novel ensemble strategy was devised by a group of researchers in order to overcome the shortcomings of past ensemble strategies, such as feature selection and low accuracy. BiLSTM or BiGRU models were combined with a CNN model to create an ensemble classifier that predicted cardiac illness with an F1 score ranging from 91% to 100% for each model. The findings suggest that ensemble frameworks are useful in dealing with the problem of forecasting with an imbalanced dataset.

3. DATA

The raw data that was used in this study was acquired through the usage of the Kaggle application. In this project, the data was organized using a comma-separated values file and columns, both of which were prepared in the Microsoft Excel programmed. Neither of the variables, which were either continuous or categorical in nature, had any values that were equal to zero at any point during their existence. Two significant flaws in the data set, which are as follows: first, the data set had an incorrect number of observations. I'll use an example to highlight how large the standard deviation was in order to give you a better idea of what I'm trying to explain. Because the dataset contained extreme values (i.e., global outliers), a global anomaly analysis was carried out on it in order to identify and classify them. For all continuous variables with a substantial standard deviation, the upper and lower 2 percent percentiles were lowered by two percentage points, allowing for an appropriate response while retaining data integrity. As well as these findings, the researchers discovered outliers such as persons who had a lower SBP than they did a greater DBP, among other things. As soon as those questionable data points were removed from the data set, the model had the ability to generate more accurate predictions than it had been able to do previously. Finally, in

Journal of Theoretical and Applied Information Technology
30th September 2024. Vol.102. No. 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

order to maintain consistency throughout the dataset, all non-categorical numeric variables were standardized between 0 and 1 in order to verify that they were uniformly distributed. Preprocessing revealed that when compared to other variables, the target variable had a disproportionately equal number of cases with and without coronary heart disease. As a result of this decision, no weighting would be applied to the target variable throughout the data analysis phase of the process. Post-processing was performed on the continuous variables (as shown in Fig. 1), and the following statistical features of the dataset were discovered.

The five continuous variables that make up the dataset are listed in the following table: Table 1. It has been decided to delete data points from the distribution of these variables that were excessive in order to improve the accuracy of the distribution. Consequently, an acceptable range of continuous variables has been established, and the data has proven to be consistent and trustworthy.

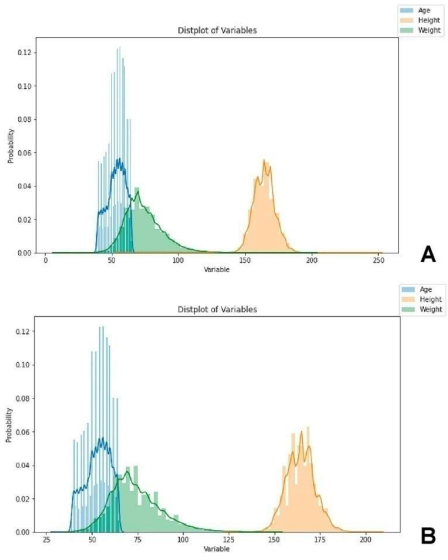


Figure 1 exhibits a graph depicting the distribution of continuous data, which depicts the distribution of continuous data. Prior to pre-processing, a distribution of variables is employed to ensure that the data is representative. B Following pre-processing, the distribution of the variables is computed.

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

The model, on the other hand, has six categorical variables, four of which are binary in nature. There are three possible values for the other two category variables cholesterol and glucose, which are 1, 2, and 3 respectively. Example: A glucose level of 3 indicates a high glucose level, whereas a glucose level of 1 indicates a low glucose level, and so on. There is a binary outcome in this dataset because it is mostly concerned with cardiovascular disease.

4.FEATURE ANALYSIS

Pearson's coefficient [30] was used to create a heat map to assess the association between characteristics and the goal variable. Clustering power was assessed by grouping comparable data points to examine the heat map's associations. Age and systolic blood pressure were mapped to the target variable and grouped together. For example, the distribution of a variable can be shown more clearly with this method.

In Fig. 2, we see a graph of two of the most important continuous variables.

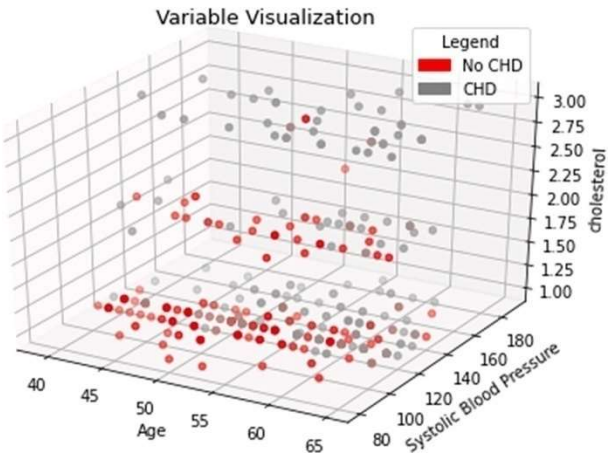


Fig. 2 depicts a three-dimensional representation of the major factors and the goal variable.

Journal of Theoretical and Applied Information Technology
30th September 2024. Vol.102. No. 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Significant coefficients are those that are greater than or equal to 0.5. For instance, it shows how the continuous variables are converted to cholesterol in the dataset. Random sampling of 400 data points was used to guarantee that the results were free of any bias. Figure 2 shows that those with heart disease had higher levels of cholesterol and blood pressure than those who were otherwise healthy. Even when employing a clustering technique to analyse data, it is not immediately clear that the age of the objective variable is linked to the aim.

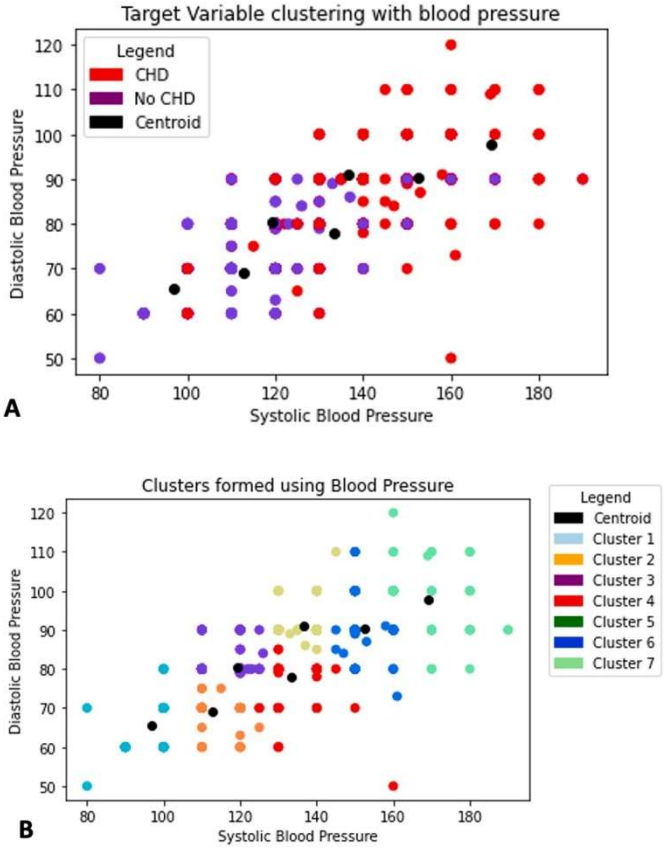
Groupings of data points that had a correlation to their aim variable were necessary to complete this assignment because it was categorised. Using centroid-based clustering, patterns were discovered in the data. The K-Means approach was used to locate the centroids in order to make data visualisation easier. The four continuous variables were plotted on a graph in order to identify clusters. These two graphs illustrate each set of variables, one with the aim variable and the other with clusters formed by employing centroids.

There are two clustering methods that can be used for blood pressure measurements, as shown in Figure 3 utilizing a computer simulation. Clustering blood pressure variables can be done effectively using this strategy, as seen in Figure 3.

The researchers found that the diastolic and systolic blood pressure measurements had the greatest Pearson and LASSO coefficients (see Fig. 4). Clusters of people with coronary heart disease and those without it are depicted in the graphs. When it comes to coronary artery disease (CAD), patients are most likely to fall into clusters 5, 6, and 7. A similar approach was used to analyse the other three continuous variables (age, height, and weight), but the results were less significant despite being less significant. K-Means and other clustering methods performed wonderfully in this classification test, according to the results.

Use a curve of best fit for each continuous variable to plot its location on a graph. Polynomial or curvilinear relationships with age were seen for several variables; however, not all of these relationships were statistically significant. Compared to the general population, patients with coronary heart disease have higher scores on all variables. Logistic classification can be a useful tool for categorising information in this instance.

Calculating Z-values and p-values for each variable (categorical as well as continuous) was necessary to bring the data analysis and quantified feature evaluation process to a close, as well as to evaluate the data set's quantified features. This might work for logistic regression models, but not for the clustering-based models, which would require a different approach. As a result of the research, it was discovered that gender had a Z-value of —0.655 and a P > |z| of 0.512, as the study's conclusion. As a result, there was no discrimination based on gender.



In Fig. 3, you can see clusters of blood pressure readings. Clusters of data points appeared. B Data points that are correlated with the goal variable.

In logistic regression, this feature is an input. Other variables had Z values that ranged from 4.21 to 60.68. (Systolic BP).

5.THE EXPERIMENTAL SETUP AND THE RESULTING DATA

Once feature analysis was done, the Least Absolute Shrinkage and Selection Operator (LASSO) was used to select the best features. Inclusion of features that aren't relevant to the job at hand may have a negative impact on model classification performance. LASSO relies on feature coefficients' absolute values being updated. Features with lower coefficients are deleted from the dataset, while those with higher values are kept.

Classifiers based on the original dataset were used to model the analyzed data.

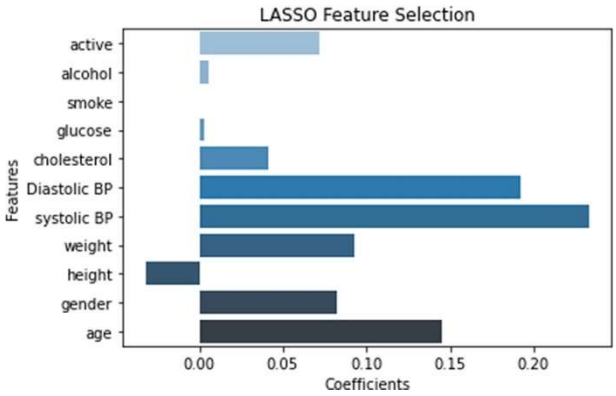


Figure 4: LASSO feature selection graph findings

determine how well they are working. Because the coefficients of alcohol, smoking, and glucose were less than 0.01, they were not included in the processed dataset. In order to train the model, 75 percent of the data were used, while 25 percent were retained for testing and analyzing performance measures. For the sake of unpredictability and not over fitting, the data points were also randomized in a database.

Using K-Fold validation, the findings of the models were tested. Ten K-Folds were utilised in this experiment. As a result, the models' performance metrics were averaged together. Finally, critical hyper-parameters, such as verbosity, iterations, and leave-nodes, were optimised using stacked loops that varied over enormous values. It was then used to discover the optimal combination of hyper-parameters for each model tested. Python was used to develop the algorithm, while sci-kit learn was utilised to model the dataset.

Classifiers of this type have been used in the past.

Table 2 shows that basic classifiers have a similar accuracy of 72.3%. Its AUC (Average Utility Coefficient) was 73.93 percent, and its accuracy was 73.93 percent (see Fig. 5).

At 73% success rate, D1 was the most accurate base classifier. It had a low recall score since it had a high AUC score. As a result, it was expected that coronary heart disease patients would be incorrectly classified. In order to improve upon the D2 decision tree's 71.4 percent accuracy, a modified version (D2) was created. In contrast, it scored 72.6 percent recall and 71.7 percent precision. There were more leaf nodes per branch in this new decision tree than in the previous one.

Model	Accuracy	Recall	Precison	10-k-fold SSD	FI score	AUC
Logistic	71.5	68.8	74.6	0.59	71.1	0.73
Classification K-NN	72.6	67.6	74.3	.044	68.8	0.72
RS-Decision Tree	74.0	64.5	78.8	0.59	69.7	0.73
J48 –SVC	73.2	63.9	75.7	0.48	68.1	0.74
Guass –Naive Bayes	72.4	62.2	77.3	0.85	68.9	0.71
Neural Network-MLP	74.9	68.2	76.2	0.63	73.0	0.74

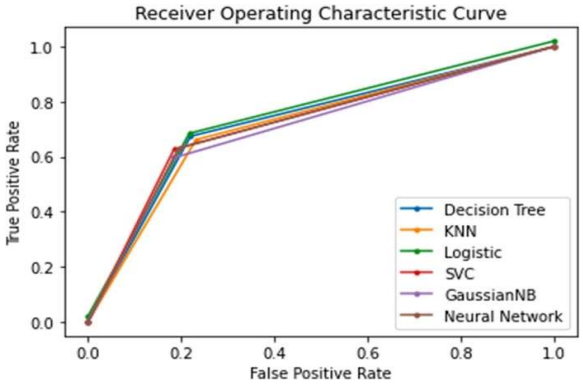


Figure 5:Base classifier receiving operator curves are shown

A dense neural network is a type of learning system used in supervised learning algorithms that adjusts its weights and biases during training in order to provide the best possible combinations. The MLP was able to arrive at the global optimal set of weights and biases for the job at hand using a fully connected neuron topology and a slow learning rate.

The proposed architecture is as follows:

- 128 neurons in the dense input layer (input dim 11). Batch Normalization and Batch Dropout are two terms that are used in batch processing (0.6)
- Hidden layer has 256 neurons and a sigmoid pattern of activity.
- Batch Normalization and Batch Dropout are two terms that are used in batch processing (0.3)
- A dense hidden layer with 256 neurons and activation 'SoftMax' is used in this model. Batch Normalization and Batch Dropout are two terms that are used in batch processing (0.15)
- Hidden layer has 256 neurons and a sigmoid pattern of activity. Normalization of many batches

- Dense Output layer, activation'relu' (output dim 2), and a dense Output layer Adam is an adaptive, dynamic learning optimizer with a learning rate of 0.01. Loss function (sometimes known as the loss function): Categorical cross entropy (also known as categorical cross entropy) is a type of entropy that may be classified into categories.

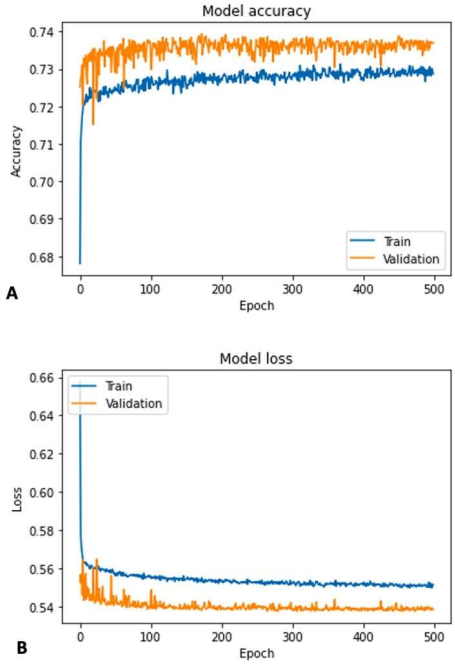


Figure 6 shows the results of the neural network evaluation. An epoch-based network accuracy measurement. B Model failure occurs in a variety of epochs

The neural network was trained for about 500 epochs to get the results shown in Fig. 6. In testing, the multi-layer perceptron-based neural network had the best accuracy of 73.9 percent and the highest F1 score of 72.0 percent. After 172 epochs, the loss function had dropped to a global minima of 0.5380 in order to get the score we needed to proceed. What you need to know about working in teams. By combining multiple heterozygous or homozygous models into a single model, ensemble techniques aim to reduce model variation. There are numerous weak classifiers which can be combined to build a strong classifier through the iterative process of boosting. Other than that, it is a homogenous technique in which classifiers for each subset of the data are fitted to assist aggregate their performance. In addition, random forests, a bagging approach that fits various subsets of data over several decision trees, will be examined. Each ensemble model's number of estimators was tweaked until the optimal number was found.

5.1 Boosting

It is a homogenous strategy in which the base classifier is trained on subsets of data in order to build many models of reasonable performance, which is referred to as boosting. The data points that were incorrectly categorized are then separated into subsets and fitted to the next model. As a result, the variance of the model is reduced by integrating several weak learners and applying a cost function to the process. The default estimator, i.e., the tree algorithm, CART, was employed as the basis estimator. In order to find the optimum number of important hyperparameters, such as estimators and the number of times the model is boosted, an iterative procedure was used to vary these values. After that, the optimal combination of hyper-parameters was selected using grid searching techniques.

The graph in Figure 7 shows that the Gradient boosting technique proved to be the most effective of the three boosted models that were assessed, while the ADA booster was shown to be the least effective. Furthermore, until there were 150 estimators, the number of estimators was directly proportional to the accuracy of the estimates. With estimators greater than 150, the accuracy of the models, with the exception of the XGB Booster, rapidly dropped over time (see Fig. 7)

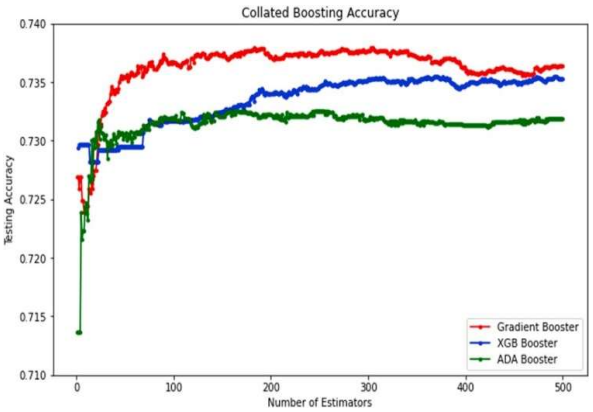


Fig. 7: Increasing the accuracy of models by using a greater number of estimators

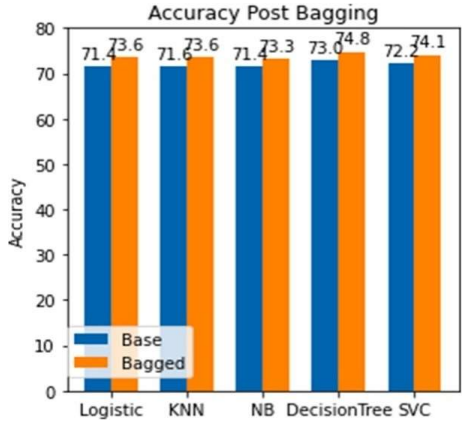


Figure 8 .Comparison of accuracy between bagged and basic models

5.2 Bagging

An aggregated predictor can only be generated by combining multiple versions of the same predictor. As a result of plurality voting, the aggregate is able to take the average of the many predictions. This improves the performance of a classifier by allowing it to run several homogeneous models in parallel and averaging their results. To determine which model was which, each model was bagged and the cross-section was compared to the original base model to determine which was which.

The use of the bagging ensemble technique proved to be advantageous, with each model's accuracy increasing by at least 1.8 percent as a result. (See Fig. 8 for an example.) In addition, the positive component increased the recall and precision ratings by a significant margin. The overall performance of the model is improved as a result of the reduction in the number of false-positive and false-negative detection rates.

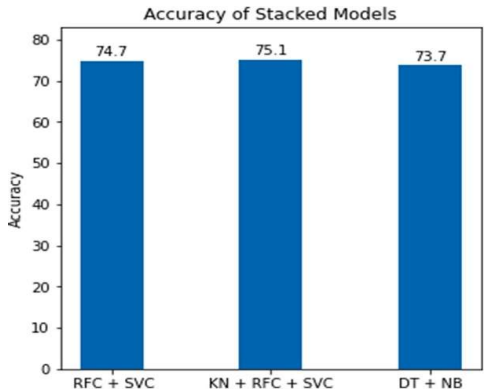


Figure 9 shows the accuracy of the best stacking models.

The statistical evaluation of bagged and boosted models is shown in Table 3.

Model	Accuracy	Recall	Precision	F1	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

The computation time is greatly reduced as compared to a single decision tree. The random forests were trimmed and the number of base estimators, or the number of trees, was changed until the optimal number was discovered. The optimal number of trees was 147, resulting in a precision of 74.42% based on the data.

Stacking

Stalling is a powerful model-building strategy that combines weak learners from many ensembles to construct a powerful model. A number of layers are generated, with each model passing its results to the one above it in the hierarchy. The ultimate judgement is made by the highest layer, whereas the inputs from the original dataset are received by the bottommost layer of the hierarchy. The majority vote is the meta classifier that is used to combine the results of the several classifiers. The binary logistic classifier was chosen as the final model since it was found to be the most successful at the topmost layer, or final model. The following table lists the base classifiers that are available for stacking.

Models such as Naïve Bayes ,Decision Tree, KNN, Random Forest and SVC are examples of classification models.

In order to get the best possible layered modelling, backtracking was employed to generate each subset of models.

The maximum performance was attained (75.1 percent accuracy) when KNN, random forest classifier, and support vector machine classifiers were stacked with logistic regression as the meta-classifier (see Fig. 9).

Proposed Model: Stacking Algorithm

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

6.DISCUSSION

Classical classifiers are the subject of most current research. Only boosting and random-forest classifiers, both of which are based on decision trees, have been studied for diagnosing coronary heart disease using ensemble approaches. In this study, it was found that stacking and bagging are more efficient and trustworthy than the other methods that were tested. Using the Cleveland dataset, researchers found that a random-forest classifier and a decision tree were the best options for classifying the data. A 74.4 percent and a 74.8 percent accuracy rate in our data showed that the random forest classifier and the decision tree to be the most reliable methods for our investigation. Previous studies demonstrate that bagged models are more effective than their traditional counterparts [34] based on this research. Many studies have shown that boosted models and random forests outperform basic classifiers in the prediction of coronary heart disease [35–37]. All measures of performance were better for the boosted models than the original classifiers. This project employed many approaches such as AdaBoost, Gradient Enhancement, and XGBoost..

The purpose of this study is to investigate whether the use of stacking may be used as an alternative to conventional methods of anticipating coronary heart disease in individuals. When stacked models were compared to base-classifiers and other ensemble processes, it was shown that stacking models was the most accurate strategy. The effectiveness of this alternate method of coronary heart disease prediction has not been thoroughly investigated in prior study. However, despite the fact that earlier research have developed models with higher accuracy, the dataset utilised in this work is far larger than the datasets used in the previous studies. Several of the earlier models are no longer applicable when dealing with real-world information.

7.CONCLUSION

Machine learning is used in this study to predict coronary heart disease, specifically. Second, the binary logistic classification was based on a thorough analysis of the data and its patterns and significant properties. For appropriate feature selection from the dataset, the statistical method in conjunction with k-nearest neighbors was crucial. It was found that the models studied could obtain a maximum accuracy of 75%. On average, the accuracy of the basic models evaluated was 71.92 percent, but the accuracy of the neural network approached 73.97 percent. With the base models, it was found that the ensemble techniques of bagging, boosting and stacking were useful.

It was found that, on average, the accuracy of the bagged models was increased by 1.9 percent, culminating in a 73.82 percent improvement in accuracy. This technique was the most effective in terms of boosting and had an accuracy rate of 73.99% when optimised for precision. Consistent results from the models were demonstrated by the K-Folds Cross-Validation, which found that accuracy ranged from 0.3% to 0.6% and standard deviations from 0.3% to 0.6%. Most effective ensemble technique was stacked with heterozygous models, which achieved a 75.1 percent accuracy rate in the experiment. With the help of logistic regression, the KNN classifier, a random forest classifier, and a support vector machine (SVM) were stacked.

Finding errors in the models using other statistical methods was less conclusive. With an aggregate accuracy rate of 76.1 percent, each model performed well. The area under the Receiving Operator Curves dropped as a result of a lower recall score. As a whole, each model earned an overall recall rate of 66.8 percent. In the future, we'll test the model's predictions against actual laboratory data to determine if they hold up. Ensemble neural networks, for example, will be investigated in the future as well. Ensemble techniques, such as boosting, bagging, or stacking, were initially considered.

REFERENCES:

[1]SchmidtH.Chronicdiseasepreventionandhealthpromotion.2016,April13. Retrieved from,https://www.ncbi.nlm.nih.gov/books/NBK435779

[2]GonsalvesAH,ThabtahF,MohammadRM,SinghG.Predictionofcoronaryheart disease using machine learning. Proceedings of the 2019 3rd international conference on deep learning. Technologies - ICDLT 2019. https://doi.org/ 10.1145/3342999.3343015.

[3] Latha CB, Jeeva SC. Improving the accuracy of prediction of heart disease risk basedonensembleclassificationtechniques.July02,https://www.sciencedirect.com/science/article/pii/S235291481830217X; 2019. Retrieved from.

[4]MuhammadYar,TahirMuhammad,HayatMaqsood,ChongKilTo.Earlyandaccurate detection and diagnosis of heart disease using intelligent computationalmodel. Sci Rep 2020;10(1):1–17.

[5] 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.14.https://doi.org/10.1109/ICSSE.2016.7551594.

[6] Dubey VK, Saxena AK. Hybrid classification model of correlation-based feature selectionandsupportvectormachine,"2016IEEEInternationalConferenceon CurrentTrendsinAdvancedComputing. Bangalore:ICCTAC);2016.p.16.https://doi.org/10.1109/ICCTAC.2016.7567338.

[7] Nilashi Mehrbakhsh, Ahmadi Hossein, Azizah Abdul Manaf, Rashid Tarik A, Samad Sarminah, Shahmoradi Leila, Aljojo Nahla, Akbari Elnaz. Coronary heart diseasediagnosisthroughself-organizingmapandfuzzysupportvectormachine
with incremental updates. Int J Fuzzy Syst 2020:1–13.

[8] Dutta Aniruddha, Batabyal Tamal, Basu Meheli, Scott T. Acton. "An efficient convolutional neural network for coronary heart disease prediction. Expert SystAppl2020;159:113408.

[9] Samuel OW, Asogbon GM, Sangaiah AK, Fang P, Li G. An integrated decision supportsystembasedonANNandFuzzy_AHPforheartfailureriskprediction.Expert Syst Appl 2017;68:163–72.

[10]SinghN,etal."Heartdiseasepredictionsystemusinghybridtechniqueofdata" mining algorithms. International Journal of Advance Research, Ideas and Innovations in Technology 2018;4(2):982–7.

[11]Nourmohammadi-KhiarakJ,Feizi-DerakhshiM,BehrouziK,etal.Newhybrid 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.

[12] Baccouche Asma, Garcia-Zapirain Begonya, Cristian Castillo Olea, Adel Elmaghraby.Ensembledeeplearningmodelsforheartdiseaseclassification:acase study from Mexico. Information 2020;11(4):207.

[13] Jain Mrs, Gupta Prof. A review and analysis of centroid estimation in k-means algorithm.IJARCCE2018;7:426.https://doi.org/10.17148/IJARCCE.2018.789.

[14] 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 congestiveheartfailurebyusingspectralclustering,"2015IEEE28thinternational symposiumoncomputerbasedmedicalsystems.SaoCarlos;2015.p.1836.https://doi.org/10.1109/CBMS.2015.71.