RESEARCH ARTICLE OPEN ACCESS

Prediction of Cardiovascular Disease with Minimal Data Using Machine Learning Algorithms, Relief, and LASSO Feature Selection

Dr K Sailaja, MCA, M.Tech, M.Phil, Ph.d [1], V Poornachandra Reddy [2]

[1] Professor, Department of Master of Computer Applications
[2] Student, Department of Master of Computer Applications
[1] , [2] Chadalawada Ramanamma Engineering College (Autonomous)

ABSTRACT

Infectious illnesses Coronary artery diseases are one of the most common causes of death worldwide. Death rates from cardiovascular diseases (CVDs) might be lowered if their onset could be averted or their effects minimised by early detection. The use of machine learning algorithms to determine potential dangers is an exciting new direction. We would want to offer a model that uses many approaches to cardiovascular disease prediction. To ensure that the suggested model is well-trained, we present a number of techniques for pre-processing and data transformation. We analysed data from the University of California, Irvine's Heart Disease dataset. Results are shown independently so that comparisons may be made. Using RFBM and Relief feature selection approaches, our suggested model achieved the maximum accuracy, as shown by our study of the results.

Keywords: - Relief Feature Selection, Decision Tree Bagging Method, Random Forest Bagging Method, K-Nearest Neighbors Bagging Method, AdaBoost Boosting Method.

I. INTRODUCTION

Cardiovascular disease has long been considered the deadliest and most debilitating condition afflicting human beings. The rising prevalence of cardiovascular illnesses is a major danger to and financial burden for healthcare systems across the globe. Although cardiovascular disease is more common in males than in women, especially in middle age and old age, it also affects youngsters. The World Health Organization reports that heart disease is responsible for one in three deaths worldwide. About 17.9 million individuals each year lose their lives to CVDs, with a greater incidence in Asia. According to the European Society of Cardiology (ESC), every year an additional 3.6 million persons are diagnosed with cardiovascular disease. About 3% of the overall health care expenditure is spent on treating heart disease, despite the fact that half of all patients diagnosed with heart disease die within only 1-2 years. Multiple diagnostic procedures are needed to anticipate cardiovascular disease. False diagnoses might occur due to medical staff's lack of competence. It is not always easy to make an early diagnosis. In underdeveloped nations, where a shortage of qualified medical personnel, testing equipment, and other resources makes it difficult to diagnose and care for individuals with heart issues, surgical treatment of heart disease presents unique challenges. An precise assessment of cardiac failure risk would aid in avoiding fatal heart attacks and increasing patient security. When given adequate training data, machine learning algorithms can successfully diagnose diseases. It is possible to evaluate different prediction models for heart disease using publicly accessible information. As a result of the advent of machine learning and artificial intelligence, scientists are now able to use the vast amounts of data at their disposal to create the most accurate prediction models possible. Recent research examining cardiac problems in both adults and children have highlighted the importance of lowering CVD-related mortality. Inconsistent and redundant clinical datasets highlight the importance of pre-processing. It's crucial to choose the relevant characteristics that may serve as risk variables in prediction models. Accurate prediction models need careful consideration when choosing the characteristics to include and the machine learning techniques to use. High frequency in most populations, independent influence on heart disease risk, and controllability or treatability are three criteria against which risk variables should be evaluated.

When modelling the predictors for CVD, researchers have included a variety of different risk variables and characteristics. Multiple studies have used features such as age, sex, chest pain (cp), fasting blood sugar (FBS) - elevated FBS is linked to Diabetes, resting electrocardiographic results (Restecg), exercise-

induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope, the number of major vessels coloured by fluoroscopy (ca), heart status (thal), maximum heart rate achieved (thalach), poor diet, and family history to develop CVD prediction models.

For the forecast to be accurate and dependable, recent research have shown that at least 14 characteristics are required.

Combining these variables with the right machine learning algorithms to generate an accurate forecast of heart disease is currently proving challenging for researchers. When trained on appropriate datasets, machine learning algorithms perform at their peak. As the accuracy of the prediction is dependent on the similarity between the training and test data, feature selection methods like data mining, Relief selection, and LASSO may be used to better prepare the data for the algorithms. Classifiers and hybrid models may then be used to provide predictions about the likelihood of illness incidence once the relevant characteristics have been picked. Classifiers and hybrid models have been developed using a variety of methods. Limited medical datasets, feature selection, ML algorithm implementations, and a lack of in-depth research are only few of the problems that may prohibit effective prediction of heart disease. The goals of our study

Prediction. Several different public data sources are used. For better forecasting results, the ensemble approach was used in Latha and Jeeva's investigation. The performance for risk detection of heart disease was deemed good after using bagging and boosting approaches to improve the accuracy of weak classifiers. In their research, the hybrid model was developed with the help of Naive Bayes, Bayes NetMultilayer Perceptron, Partial Averaging Recurrent Neural Networks (PART), and Random Forest (RF) classifiers. The created model was able to reach an accuracy of 85.48 percent. There have been recent experiments using the UCI Heart Disease dataset to compare traditional and machine learning approaches, such as RF, Support Vector Machine (SVM), and learning models. Its precision

was enhanced by using several classifiers and a voting-based strategy. In spite of their weak classification abilities, the study's subjects saw significant gains. NK. Kumar and Sikamani employed a variety of machine learning classification methods to make predictions about chronic illness. When used to CVD prediction, the Hoeffding classifier showed an accuracy of 88.56 percent in their research. Many different learning algorithms and ensemble methods were used by Ashraf et al. for their predictions: Bayes Net, J48, KNN, multilayer perceptron, Naive Bayes, random tree, and random forest. For example, J48 was the most accurate of them at 70.77 percent. The team then resorted to cutting-edge methods, ultimately resulting in an 80% accuracy rate for KERAS. For the purpose of predicting the development of Cardiovascular disease, it was suggested to use a multitask (MT) recurrent neural network, which makes use of the attention mechanism. An improvement in Area under Curve (AUC) of 2-6% is achieved for the suggested model. Amin et al. used a variety of machine learning models (k-NN, DT, NB, LR, SVM, Neural Network, and a hybrid of voting with NB and LR) to determine which was best at predicting crucial risk indicators. The research concluded that an accuracy of 87.41% could be attained using the hybrid model in conjunction with the chosen characteristics. Saqlain et almethodology .'s included the use of the SVM classification model and the mean Fisher score feature selection algorithm (MFSFSA). The desired subset of features was produced by SVM, and a validation procedure was utilised to determine MCC. A higher-than-average Fisher score was used to pick the characteristics. The combined MFSFSA and SVM had an accuracy of 81.19 percent, a sensitivity of 72.92 percent, and a specificity of 88.68 percent. Mienye et al. offer a heart disease prediction model by using a mixture of a mean-based splitting approach, a classification tree, and a regression tree to randomly divide the dataset into smaller groups. After that, a consistent ensemble was produced using a weighted classifier ensemble based on accuracy, and it achieved 93% and 91% accuracy in classification, respectively, on the Cleveland and Framingham test sets. The research by Tama et al. proposes a two-stage ensemble-based CHD detection methodology. For this purpose, we used three distinct ensemble learners: random forest, gradient boosting machine, and extreme gradientboosting machine. The suggested model has better accuracy (98.13%), F1 (96.6%), and AUC (98.7%) than previously available techniques for detecting CHD.

In their publication, Mohan et al. proposed an unique prediction model that makes use of a wide variety of feature combinations and established categorization methods. The proposed HRFLM uses an ANN trained using back propagation and 13 clinical characteristics as input, and it takes into account DT, NN, SVM, and KNN when employing data mining techniques. Using SVM helped improve the reliability of disease forecasts. Combining the cutting-edge Vote technique with a mixed LR and

II. RELATEDWORKS

Due to the increased precision and efficiency of predictions, the use of AI and machine learning algorithms has exploded in recent years [25]. In order to create and choose the most accurate and efficient models, this field of study is crucial [26]. A potential method for illness prediction [27] is the use of hybrid models, which combine several machine learning models with information systems (major components). Several different public data sources are used.

Latha and Jeeva [28] used an ensemble approach to boost the precision of their predictions. The performance for risk detection of heart disease was deemed good with the use of bagging and boosting to improve the accuracy of weak classifiers. Naive Bayes, Bayes Net, C 4.5, Multilayer Perceptron, PART, and Random Forest (RF) classifiers were utilised to construct the hybrid model based on majority voting. The created model was able to reach an accuracy of 85.48 percent. Recently [29] the UCI Heart Disease dataset was used to evaluate both machine learning and traditional approaches, such as RF, Support Vector Machine (SVM), and learning models. Combining several classifiers with a voting-based model increased accuracy. Research indicated that even the weakest classifiers improved by 2.1%. Research by NK. Kumar and Sikamani [30] uses a variety of machine learning categorization methods to make predictions about chronic illness.

They found that the Hoeffding classifier was 88.56 percent accurate in predicting CVD in their investigation.

For their predictions, Ashraf et al. [15] employed a variety of different learning algorithms and ensemble methods, including Bayes Net, J48, KNN, multilayer perceptron, Naive Bayes, random tree, and random forest. For example, J48 was the most accurate of them at 70.77 percent. The team then resorted to cutting-edge methods, ultimately resulting in an 80% accuracy rate for KERAS. The development of cardiovascular illness may be predicted using a multitask (MT) recurrent neural network, which takes use of the attention mechanism [16]. An improvement in Area under Curve (AUC) of 2-6% is achieved for the suggested model.

Amin et al. [12] used machine learning models (k-NN, DT, NB, LR, SVM, Neural Network, and a hybrid of voting with NB and LR) to the selected significant risk indicators and compared their performance.

Using the hybrid model and the predetermined characteristics, they were able to attain an accuracy of 87.41%, as shown by their research. Saqlain et al. [31] suggested a method that combines the SVM classification model with the mean Fisher score feature selection algorithm (MFSFSA).

The desired subset of features was produced by SVM, and a validation procedure was utilised to determine MCC.

A higher-than-average Fisher score was used to pick the characteristics. A combined MFSFSA and SVM achieved an 81.19 percent success rate with a sensitivity of 72.92 percent and a specificity of 88.6 percent.

Mienye et al. [22] propose a heart disease prediction model by using a mean-based splitting method, a classification tree, and a regression tree to randomly partition the dataset into smaller subsets; this is followed by the generation of a homogeneous ensemble using a weighted classifier ensemble, which achieves 93% and 91% classification accuracies on the Cleveland and Framingham test sets, respectively. The research of Tama et al. [24] proposes a two-stage ensemble-based CHD detection methodology. Random forest, a gradient boosting machine, and an extreme gradient boosting machine were the three ensemble learners used. The suggested model has higher accuracy (98.13%), F1 (96.6%), and AUC (98.5%) than previously available CHD detection approaches.

Mohan et al. [32] proposed a unique prediction model that makes use of a wide variety of feature combinations and established categorization methods. The suggested HRFLM uses an ANN trained with backpropagation and fed with data consisting of 13 clinical characteristics. While implementing the data mining techniques, DT, NN, SVM, and KNN were taken into account. Using SVM helped improve the reliability of illness forecasts. It was suggested to use the unique technique Vote in tandem with a hybrid strategy including LR and NB. The HRFLM technique achieved an accuracy of 88.7 percent.

Comprehensive risk modelling for death prediction in heart failure was developed using an improved random survival forest (iRSF) [33].

Using the new split rule and the stop criterion, iRSF was able to distinguish between survivors and non-survivors. The 32 risk variables used in developing predictors comprised patient demographics, clinical, laboratory information, and medicines. In addition, data mining has been used for the detection of cardiovascular issues [34]. Data mining methods such as the Decision Tree, Bayesian classifier, neural network, Association law, SVM, and KNN were utilised to diagnose heart disease.

The accuracy achieved using SVM was 99.3 percent.

Several machine learning classifiers have been used in research on patient survival prediction [35].

Ranking features related to the most important risk variables and contrasting the results of conventional biostatistical tests with those of the given machine learning methods were done. Serum creatinine and ejection fraction were shown to be the two most important factors for precise forecasting. Using the AL Algorithm, [36] we were able to create a model for CVD identification. Four methods were used for the initial dataset creation and subsequent examination. Accuracy for DT and RF techniques was 99.83%, while accuracies for SVM and KNN were 85.32% and 84.49%, respectively. Another research [37] used deep neural networks to successfully predict congestive heart failure (CHF) by examining HRV, filling a need in the field. The suggested system has a 99.85% success rate.

The research of Yadav and Pal [3] relied on data kept in the UCI repository.

There are 14 distinguishing characteristics in this data collection. Four tree-based classification algorithms—M5P, random Tree, Reduced Error Pruning, and the Random forest ensemble method—performed the classification. This study used three feature-based algorithms: the Pearson Correlation, Recursive Features Elimination, and Lasso Regularization. The strategies were then evaluated for their precision and accuracy, with the final approach yielding the best results. In recent research [38], Gupta et al. used RF-based MLA and factor analysis of mixed data (FAMD) to create a framework for artificial intelligence. The FAMD was utilised to identify the pertinent aspects for RF's disease prediction purposes. The suggested strategy was successful in predicting outcomes with a 93.44% accuracy, 89.28% sensitivity, and 96.96% specificity.

III. PROPOSED SYSTEM ARCHITECTURE

In our proposed model, ten features have been evaluated to make this comparison more unique. The introduced algorithms were conducted based on the all features, Relief selected features the obtained outcomes were compared to other works to show the percentage of improvement, while decrease in performance also noted in one occasion (RFBM, DTBM, KNNBM, ABBM, GBBM). The highest increment was noticed for AB approach as opposed to previous works which was about percentage improvement were calculated for 13 attributes.Cardiovascular disease is used to determine whether or not a patient is at risk of having a heart attack.

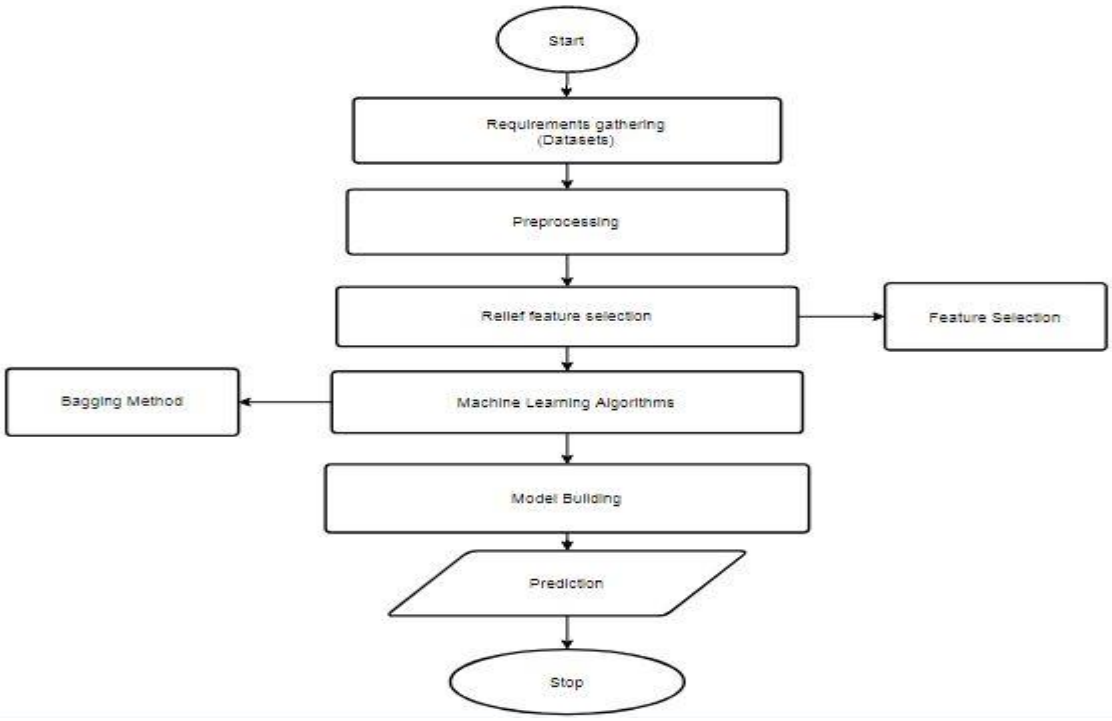

Fig.1 Proposed System

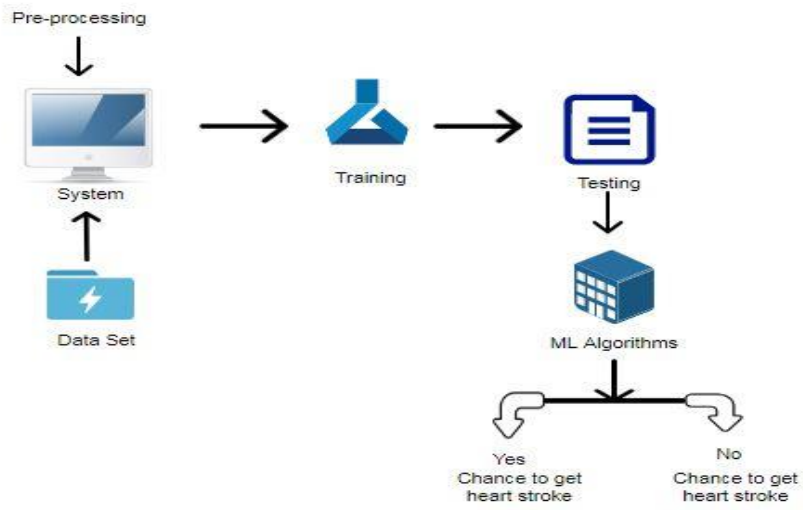

Fig.2 Methodology

1. Decision Tree Bagging Method:

Decision Tree is a Supervised Machine Learning approach to solve classification and regression problems by continuously splitting data based on a certain parameter. The decisions are in the leaves and the data is split in the nodes. In Classification Tree the decision variable is categorical (outcome in the form of Yes/No) and in Regression tree the decision variable is continuous. Decision Tree has the following advantages: it is suitable for regression as well as classification problem, ease in interpretation, ease of handling categorical and quantitative values, capable of filling missing values in attributes with the most probable value, high performance due to efficiency of tree traversal algorithm.

Decision Tree might encounter the problem of over-fitting for which Random Forest is the solution which is based on ensemble modelling approach. Disadvantages of decision tree is that it can be unstable, it may be difficult to control size of tree, it may be prone to sampling error and it gives a locally optimal solution- not globally optimal solution. Decision Trees can be used in applications like predicting future use of library books and tumour prognosis problems.

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm.

Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram.

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.** By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are Information Gain Gini Index.

Bootstrap Aggregation is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosenrandomly *with replacement.* Now, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree.

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement meaning that the individual data points can be chosen more than once. After several data samples are generated, these weak models are then trained independently, and depending on the type of task regression or classification, for example the average or majority of those predictions yield a more accurate estimate. As a note, the random forest algorithm is considered an extension of the bagging method, using both bagging and feature randomness to create an uncorrelated forest of decision trees.

Bagging and boosting are two main types of ensemble learning methods. As highlighted in this study the main difference between these learning methods is the way in which they are trained. In bagging, weak learners are trained in parallel, but in boosting, they learn sequentially. This means that a series of models are constructed and with each new model iteration, the weights of the misclassified data in the previous model are increased. This redistribution of weights helps the algorithm identify the parameters that it needs to focus on to improve its performance. AdaBoost, which stands for “adaptative boosting algorithm,” is one of the most popular boosting algorithms

2. Random Forest Bagging Method:

Random forest is an extension over bagging. It takes one extra step where in addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees. When you have many random trees. It’s called Random Forest Suppose there are N observations and M features in training data set. First, a sample from training data set is taken randomly with replacement. A subset of M features are selected randomly and whichever feature gives the best split is used to split the node iteratively The tree is grown to the largest Above steps are repeated and prediction is given based on the aggregation of predictions from n number of trees Handles higher dimensionality data very well. Handles missing values and maintains accuracy for missing data.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

The random forest algorithm is actually a bagging algorithm: also here, we draw random bootstrap samples from your training set. However, in addition to the bootstrap samples, we also draw random subsets of features for training the individual trees; in bagging, we provide each tree with the full set of features. Due to the random feature selection, the trees are more independent of each other compared to regular bagging, which often results in better predictive performance (due to better variance-bias trade-offs), and I’d say that it’s also faster than bagging, because each tree learns only from a subset of features.

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Suppose we have 1000 observations in the complete population with 10 variables. Random forest will try to build multiple CART along with different samples and different initial variables. It will take a random sample of 100 observations and then chose 5 initial variables randomly to build a CART model. It will go on repeating the process say about 10 times and then make a final prediction on each of the observations. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

Bagging is an ensemble algorithm that fits multiple models on different subsets of a training dataset, then combines the predictions from all models. Random forest is an extension of bagging that also randomly selects subsets of features used in each data sample. Both bagging and random forests have proven effective on a wide range of different predictive modelling problems.Although effective, they are not suited to classification problems with a skewed class distribution. Nevertheless, many modifications to the algorithms have been proposed that adapt their behaviour and make them better suited to a severe class imbalance.

Bootstrap Aggregation, also known as bagging, is a powerful ensemble method that was proposed by Leo Breiman in 1994 to prevent over fitting. The concept behind bagging is to combine the predictions of

several base learners to create a more accurate output. Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees.This approach can be used with machine learning algorithms that have a high variance, such as decision trees. A separate model is trained on each bootstrap sample of data and the average output of those models used to make predictions. This technique is called bootstrap aggregation or bagging for short.

Variance means that an algorithm's performance is sensitive to the training data, with high variance suggesting that the more the training data is changed, the more the performance of the algorithm will vary.

Bootstrap Aggregation is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement.

3. K-Nearest Neighbors Bagging Method

K Nearest Neighbour (KNN) Algorithm is a classification algorithm It uses a database which is having data points grouped into several classes and the algorithm tries to classify the sample data point given to it as a classification problem. KNN does not assume any underlying data distribution and so it is called non-parametric. Advantages of KNN algorithm are the following: it is simple technique that is easily implemented. Building the model is cheap. It is extremely flexible classification scheme and well suited for Multi-modal classes. Records are with multiple class labels. Error rate is at most twice that of Bayes error rate. It can sometimes be the best method. KNN outperformed SVM for protein function prediction using expression profiles. Disadvantages of KNN are the following: classifying unknown records are relatively expensive. It requires distance computation of k-nearest neighbours.

With the growth in training set size the algorithm gets computationally intensive, Noisy / irrelevant features will result in degradation of accuracy. It is lazy learner; it computes distance over k neighbours. It does not do any generalization on the training data and keeps all of them. It handles large data sets and hence expensive calculation. Higher dimensional data will result in decline in accuracy of regions. KNN can be used in Recommendation system, in medical diagnosis of multiple diseases showing similar symptoms, credit rating using feature similarity, handwriting detection, analysis done by financial institutions before sanctioning loans, video recognition, forecasting votes for different political parties and image recognition.

A k-nearest neighbor (KNN) based bagging pruning algorithm for ensemble KNN classification is proposed in this paper. Redundant bags are discarded without reducing the performance of the ensemble classifier. Ten VCI binary classification datasets are used to evaluate the performance of the proposed pruning algorithm against single and bagging classifiers. Results show that the proposed bagging pruning improves the classification accuracies on most of the datasets with use less number of base classifiers thereby reducing computational requirements.

An experimental evaluation of Bagging K-nearest neighbor classifiers (KNN) is performed. The goal is to investigate whether varying soft methods of aggregation would yield better results than Sum and Vote. We evaluate the performance of Sum, Product, MProduct, Minimum, Maximum, Median and Vote under varying parameters. The results over different training set sizes show minor improvement due to combining using Sum and MProduct. At very small sample size no improvement is achieved from bagging KNN classifiers. While Minimum and Maximum do not improve at almost any training set size, Vote and Median showed an improvement when larger training set sizes were tested. Reducing the number of features at large training set size improved the performance of the leading fusion strategies.

4. AdaBoost Boosting Method:

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learning. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference. Let's discuss this difference in detail.

First, let us discuss how boosting works. It makes 'n' number of decision trees during the data training period. As the first decision tree/model is made, the incorrectly classified record in the first model is given priority. Only these records are sent as input for the second model. The process goes on until we specify a number of base learners we want to create. Remember, repetition of records is allowed with all boosting techniques.

This figure shows how the first model is made and errors from the first model are noted by the algorithm. The record which is incorrectly classified is used as input for the next model. This process is repeated until the specified condition is met. As you can see in the figure, there are 'n' number of models made by taking the errors from the previous model. This is how boosting works. The models 1, 2, 3… N are individual models that can be known as decision trees. All types of boosting models work on the same principle.

Since we now know the boosting principle, it will be easy to understand the AdaBoost algorithm. Let's dive into AdaBoost's working. When the random forest is used, the algorithm makes an 'n' number of trees. It makes proper trees that consist of a start node with several leaf nodes. Some trees might be bigger than others, but there is no fixed depth in a random forest. With AdaBoost, however, the algorithm only makes a node with two leaves, known as Stump.

5. Gradient Boosting Boosting Method:

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.Unlike, Adaboosting algorithm, the base estimator in the gradient boosting algorithm cannot be mentioned by us. The base estimator for the Gradient Boost algorithm is fixed and i.e. Decision Stump. Like, AdaBoost, we can tune the estimator of the gradient boosting algorithm. However, if we do not mention the value of estimator, the default value of estimator for this algorithm is 100.Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

Gradient boosting is one of the most powerful techniques for building predictive models.In this post you will discover the gradient boosting machine learning algorithm and get a gentle introduction into where it came from and how it works

The idea of boosting came out of the idea of whether a weak learner can be modified to become better. Michael Kearns articulated the goal as the "Gradient Boosting Problem" stating the goal from a practical standpoint.Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function

.

IV. RESULTS AND DISCUSSION

The output screens are shown from Fig.3 to Fig. 7

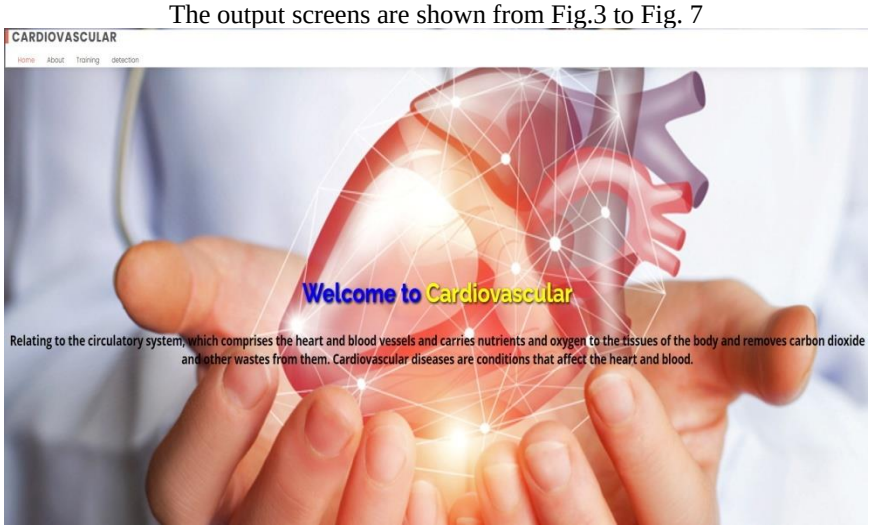

Fig.3 Home Page

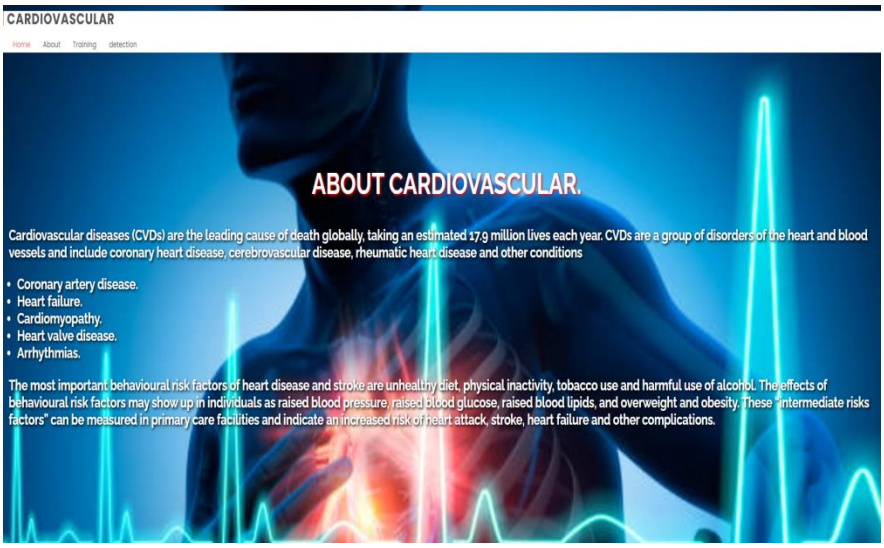

Fig. 4 About Page

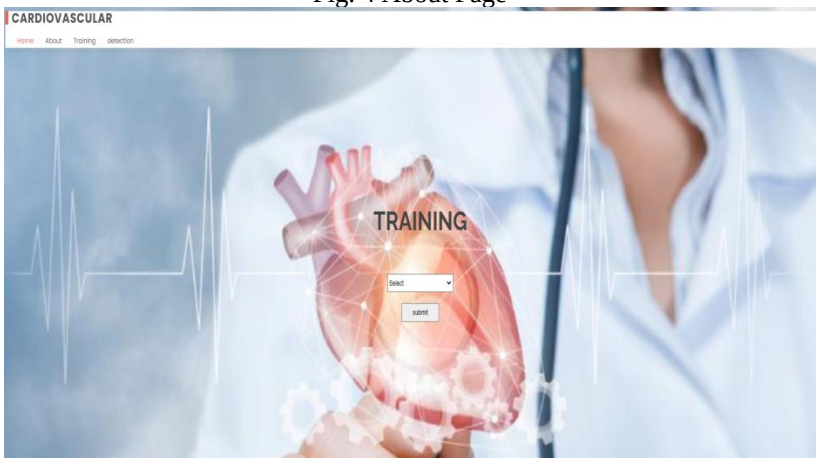

Fig.5 Training Algorithm

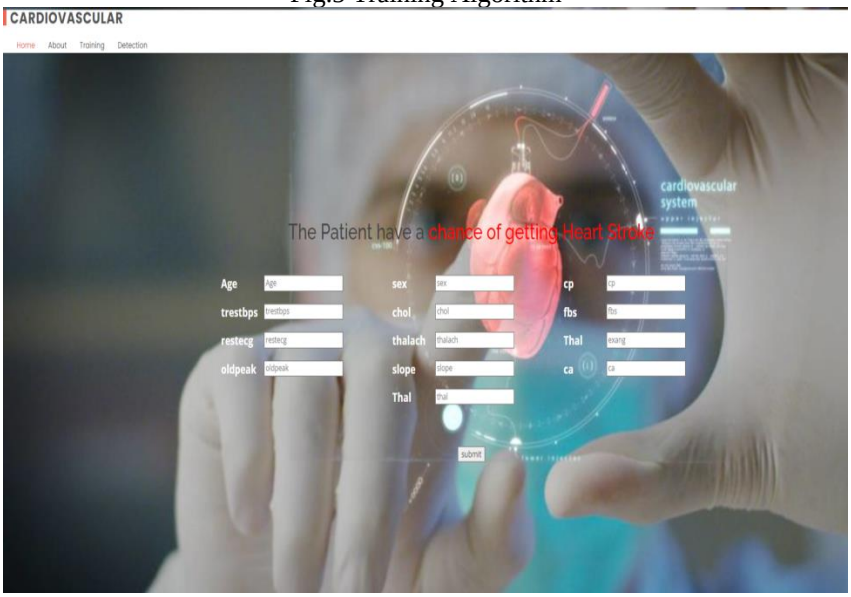

Fig.6The Patient have a chance of getting Heart Stroke

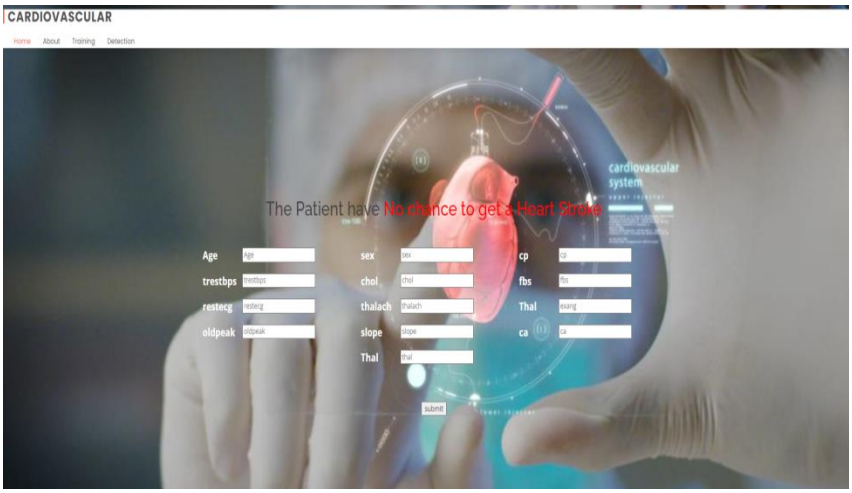

Fig.7The Patient have No chance of getting Heart Stroke

V. FUTURE SCOPE AND CONCLUSION

This study takes similar route, but with an improved and novel method and with a larger dataset for training the model. This research demonstrates that the Relief feature selection algorithm can provide a tightly correlated feature set which then can be used with several machine learning algorithms. The study has also identified that RFBM works particularly well with the high impact features and produces an accuracy, substantially higher than related work. RFBM achieved abestaccuracy with 13 features. Cardiovascular disease is used to determine whether or not a patient is at risk of having a heart attack.

REFERENCES

[1] C. Trevisan, G. Sergi, S. J. B. Maggi, and H. Dynamics, ‘‘Gender differences in brain-heart connection,’’ in Brain and Heart Dynamics. Cham, Switzerland: Springer, 2020, p. 937.

[2] M. S. Oh and M. H. Jeong, ‘‘Sex differences in cardiovascular disease risk factors among Korean adults,’’ Korean J. Med., vol. 95, no. 4, pp. 266–275, Aug. 2020.

[3] D. C. Yadav and S. Pal, ‘‘Prediction of heart disease using feature selection and random forest ensemble method,’’ Int. J. Pharmaceutical Res., vol. 12, no. 4, 2020.

[4] World Health Organization and J. Dostupno, ‘‘Cardiovascular diseases: Key facts,’’ vol. 13, no. 2016, p. 6, 2016. [Online]. Available: https:// www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases- (cvds)

[5] K. Uyar and A. Ilhan, ‘‘Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks,’’ Procedia Comput. Sci., vol. 120, pp. 588–593, Jan. 2017.

[6] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, ‘‘A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,’’ Mobile Inf. Syst., vol. 2018, pp. 1–21, Dec. 2018.

[7] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, ‘‘A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,’’ in Proc. IEEE Symp. Comput. Commun. (ISCC), Jul. 2017, pp. 204–207.

[8] J. Mourao-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, ‘‘Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data,’’ NeuroImage, vol. 28, no. 4, pp. 980–995, Dec. 2005.

[9] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ‘‘Innovative artificial neural networks-based decision support system for heart diseases diagnosis,’’ J. Intell. Learn. Syst. Appl., vol. 5, no. 3, pp. 176–183, 2013.

[10] Q. K. Al-Shayea, ‘‘Artificial neural networks in medical diagnosis,’’ Int. J. Comput. Sci., vol. 8, no. 2, pp. 150–154, 2011.

[11] F. M. J. M. Shamrat, M. A. Raihan, A. K. M. S. Rahman, I. Mahmud, and R. Akter, ‘‘An analysis on breast disease prediction using machine learning approaches,’’ Int. J. Sci. Technol. Res., vol. 9, no. 2, pp. 2450–2455, Feb. 2020.

[12] M. S. Amin, Y. K. Chiam, and K. D. Varathan, ‘‘Identification of significant features and data mining techniques in predicting heart disease,’’ Telematics Informat., vol. 36, pp. 82–93, Mar. 2019.

[13] N. Kausar, S. Palaniappan, B. B. Samir, A. Abdullah, and N. Dey, ‘‘Systematic analysis of

applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients,'' in Applications of Intelligent Optimization in Biology and Medicine. Cham, Switzerland: Springer, 2016, pp. 217–231.
[14] J. Mackay and G. A. Mensah, ''The atlas of heart disease and stroke,'' World Health Org., Geneva, Switzerland, Tech. Rep., 2004.
[15] M. Ashraf, S. M. Ahmad, N. A. Ganai, R. A. Shah, M. Zaman, S. A. Khan, and A. A. Shah, Prediction of Cardiovascular Disease Through Cutting-Edge Deep Learning Technologies: An Empirical Study Based on TENSORFLOW, PYTORCH and KERAS. Singapore: Springer, 2021, pp. 239–255.
[16] F. Andreotti, F. S. Heldt, B. Abu-Jamous, M. Li, A. Javer, O. Carr, S. Jovanovic, N. Lipunova, B. Irving, R. T. Khan, R. Dürichen, ''Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units,'' 2020, arXiv:2007.08491. [Online]. Available: https://arxiv.org/abs/2007.08491
[17] W. Wiharto, H. Kusnanto, and H. Herianto, ''Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis,'' Int. J. Electr. Comput. Eng., vol. 7, no. 2, p. 1023, Apr. 2017.
[18] A. K. Paul, P. C. Shill, M. R. I. Rabin, and M. A. H. Akhand, ''Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease,'' in Proc. 5th Int. Conf. Informat., Electron. Vis. (ICIEV), May 2016, pp. 145–150.
[19] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, ''A hybrid classification system for heart disease diagnosis based on the RFRS method,'' Comput. Math. Med., vol. 2017, pp. 1–11, Jan. 2017.
[20] D. Singh and J. S. Samagh, ''A comprehensive review of heart disease prediction using machine learning,'' J. Crit. Rev., vol. 7, no. 12, p. 2020, 2020.