International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Empirical Analysis of Cardiovascular Diseases using Machine Learning and Soft Computing Techniques

Raghavendra Kumar, Ashish Mishra, Himanshu Rathore

Abstract*: Cardiovascular diseases are a one of the most exigent issue in healthcare domain. There have been various multidisciplinary approaches proposed and applied to reduce the mortality rate. As per literature and current study machine learning and soft computing techniques are efficient and widely accepted approaches in research community. This paper identifies and compares the various techniques of machine learning using Random Forest (RF), Support Vector Machine (SVM), XGBoost and Artificial Neural Network (ANN) and uncovers the F1 score, recall, precision to predict efficient and more accurate result. The results are further compared with existing benchmark models and showed significant improvement in heart disease prediction of patient.

***Keywords*: Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN)**

I. INTRODUCTION

Cardiovascular diseases are various diseases that affect the functioning of the heart. They are the number one cause of deaths around the world [1]. Approximately 17.9 million individuals die each year from heart disease [1]. People with cardiac illness or those at elevated cardiac danger need early detection and management using counseling and medications, which is why prediction is so essential in these illnesses. We plan to achieve that by using machine learning algorithms. This is a tough task as it poses a lot of challenges such as the high risk in the event of a false negative, which could even be fatal, or not having relevant data provided for. Selecting the relevant features and working on them is very important to find good prediction models. The common specific medical attributes used include chest pain type, resting blood pressure, cholesterol levels, electrocardiograph results (resting), number of major vessels coloured by fluoroscopy, and so on. Other than these, common attributes such as age or sex are also utilized. Various algorithms were tried, including KNN, Random Forest, Decision Trees, SVM, XGBoost, Cat-Boost, LightGBM and artificial neural networks. Least amount of log loss was needed as the Stacking and ensemble modeling was eventually used to find the best models possible.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Raghavendra Kumar*, Department of Information Technology, KIET Group of Institutions, Ghaziabad, India. Email: raghavendra.dwivedi@gmail.com
Ashish Mishra, Department of Information Technology, KIET Group of Institutions, Ghaziabad, India.Email: a0505mishra@gmail.com
Himanshu Rathore, Department of Information Technology, KIET Group of Institutions, Ghaziabad, India. Email: *himanshu.1613038@kiet.edu*

II. RELATED WORK

K. Polaraju et. al [2], proposed the usage of Multiple Regression Models to predict Heart Diseases. The dataset used, the training data, was made up of 3000 cases. It had 13 characteristics, which were used as the grounds for model training. Training information accounted for 70% of the entire dataset, while the remainder, i.e.,30%, were sent to the test information. Results analysis leads us to think that, while predicting cardiovascular diseases, regression is better than most other models. Das et. al [3]employed the services of an ensemble method consisting of neural networks for diagnosing of the heart disease. It merged the posterior probabilities of various predecessor models or the expected values. The classification accuracy obtained by them using this ensemble model was 89.01%. The dataset that they used for training and testing was the Cleveland Heart Database. Anbarasi et. al [4] used almost a 2-step approach where first they used 3 classifiers. These classifiers that they tried are decision trees, Naïve Bayes and classification by clustering using thirteen attributes. In the next step, they used genetic algorithm to apply feature subselection and acquired nearly comparable precision. Their findings showed that decision tree classifier performed the best at 99.2 percent for binary classification. The Naïve Bayes classifier followed with the result of 96.5% and then the classification cluster with 88.3%. Zhang et. al [5] used SVM (Support Vector Machines) as the model for cardiovascular disease prediction. They employed PCA to find the relevant and important features and then used various kernels such as rbf and linear to work on the aforementioned features. Radial Basis Function (RBF) gave the highest classification accuracy. To discover the appropriate parameter values, the grid search method was used. In binary classification, the classification precision was 88.6364 percent. Noura Ajam [6] used artificial neural networks. Feed-forward Back propagation algorithms were used to train the model based on the specified requirements. On finding the appropriate parameters, classification accuracy reached to 88%. The number of neurons that were employed in the hidden layer were 20. ANN shows good results as it is a suitable approach to use when the amount of data is large, as in large datasets. Elshazly et. al [7] proposed a novel approach for classification, GASVM, for diagnosis of lymph diseases.

The GA part of it is just there to make the feature pool reduced. Cross validation technique used was k-fold. Different kernel features were used and performance measurements such as precision, sensitivity, area under curve (AUC), F-measurement were assessed for each feature. Final classifier that was found to be giving the most optimal results was linear, with a score of 83.1%. Dey et. al [8] used multiple models including SVM, Naive Bayes and Decision tree. They also experimented with feature selection and worked using it and without using it as well. The method used here was principal component analysis. The dataset is a binary classification problem. Final observations indicated that SVM outperformed the other two and was the best choice for the classifier. Weng et. al [9] used a similar approach to Dey et. al. where they compared various models such as random forest, gradient boosting, neural networks and logistic regression. These 4 machine learning algorithms were used to try and predict cardiovascular diseases. Grid search was used for parameter optimization. PCA was used as well and the best model based on all these parameters was found.

III. METHODOLOGY

After completing the existing research study we identified the model involves SVM, XGBoost, Random Forest algorithms as machine learning and Artificial Neural Network (ANN) as soft computing technique. These four methods form an ensemble model where averaging the prediction probabilities out and then passing that to the log loss equation in Fig 1. An ensemble model like Das et al. [3] used but other than that we are also using algorithms such as XGBoost and ANN which haven't really been used in this way in ensemble with SVM so that makes our approach unique. Also, the metric that we used for evaluation is Log Loss, which heavily penalizes large deviations in incorrect predictions. This leads to our model being different than a lot of approaches tried beforehand. Making an ensemble is the major step which led to the most improvement that we have had till now in our experiments.

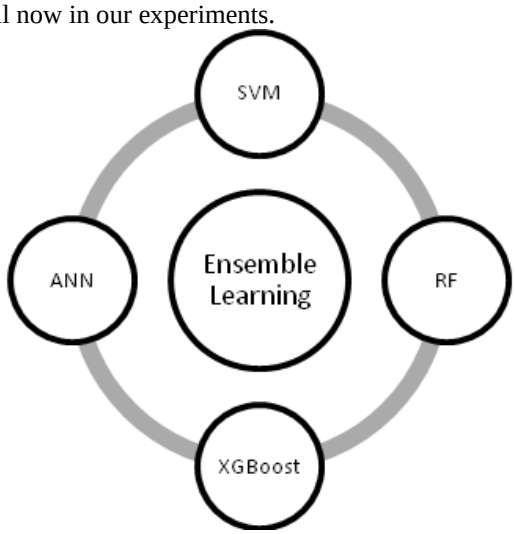

Fig-1 Ensemble Model Composition

The idea behind a support vector machine is that we have some data points that we want to label either A or B, and we can do that by plotting a line, plane or hyper plane between the two classes A and B, depending upon the problem. What an SVM does is figure out the "best" hyperplane to divide the points, which is the one that leaves the biggest margin between itself and the points. That line is called the maximum margin hyperplane, because it is a hyperplane that separates the sides while leaving the maximum margin. On trying the various kernels, RBF was found to be the most accurate one, and the one that gave the least log loss. XGBoost stands for eXtreme Gradient Boosting. XGBoost is an optimized library for boosting distributed gradients. It is intended to be flexible and extremely effective. It uses algorithms for machine learning under the framework for gradient boosting.

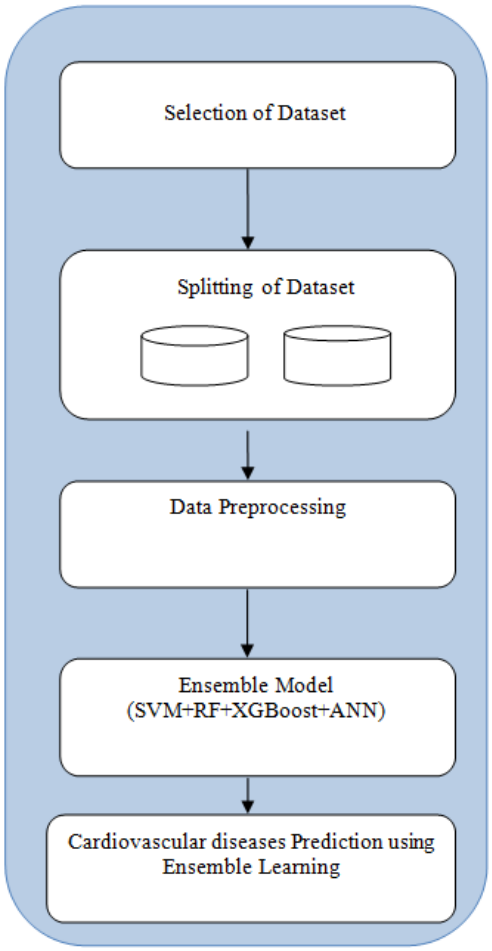

Fig 2 Flow chart for prediction model

IV. RESULT ANALYSIS AND DISCUSSION

Cleveland database is considered as training and testing dataset to predict cardiovascular decease in patient. It is popular multivariate dataset in research community used for classification in machine learning approach [11]. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The dataset used was the dataset given over at the competition on predicting heart diseases using machine learning at driven data.

The dataset has 13 essential features, and 180 instances. Patient id is a feature as well but it must be dropped. There features are not dependent on each other. The metric used for the competition we took part in is logarithmic loss.

$$Log\ loss = \sum n\ i{=}1\ [yi\ \log(y`i)+(1-yi\ )\log(1-y`i)] \quad (1)$$

where y` is the probability that y=1. The goal is to reduce log loss. Log loss provides a penalty for too deviant of predictions. The testing procedure involved us passing the data through the model, noting down the accuracy and log loss and repeating it for the various models.

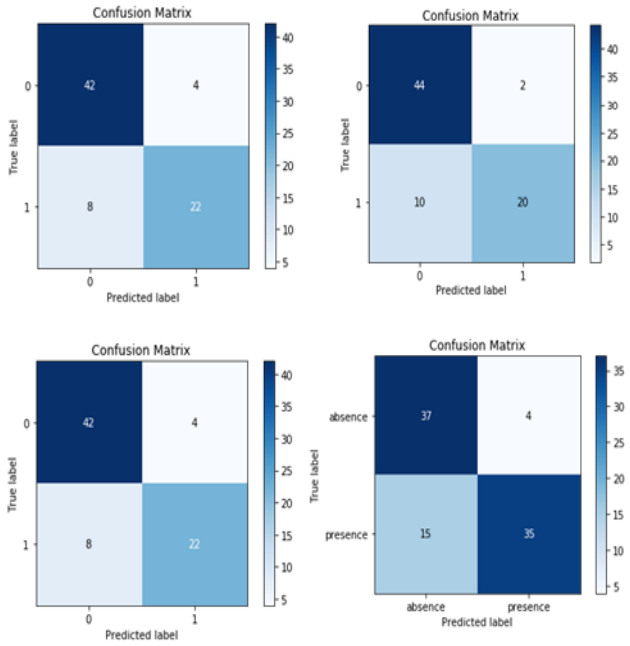

Fig 3 Confusion Matrix of Support Vector Matrix (SVM), Artificial Neural Network (ANN), Random Forest(RF) and Xgboost

We began our experimentations by working with the dataset, cleaning it up, one-hot encoding the categorical variables, normalizing with standard scaler and splitting it into 70% training and 30% testing dataset. Following that began experimentations with SVM and Neural networks at first. They were tried separately, with SVM giving the least log loss among the two at first, optimized using grid search.

After optimizing the hyper parameters of NN, we got it close to SVM but it could never cross it. So we decided to go with SVM. Performance measures are taken to estimate result of Cardiovascular diseases forecasting. It holds Recall, Precision, F1-score and accuracy.

These are considered and results are obtained with the help of confusion matrix shown in Fig 3. True Positive (TP) is the condition where the number of instances classified as true while they were actually true. False Positive (FP) is the condition where the number of instances classified as true while they were actually false.

Table-3 Comparison of results of different methods

Methods	Precision	Recall	f1-score	Support	Accuracy(%)
XGBoost	0.81	0.79	0.79	91	79.12
RF	0.84	0.84	0.84	76	84.21
SVM	0.87	0.86	0.85	80	87.21
ANN	0.89	0.88	0.88	86	89.31

False Negative (FN) is the condition here the number of instances classified as false while they were actually true. True Negative (TN) is the condition where the number of records classified as false while they were actually false.

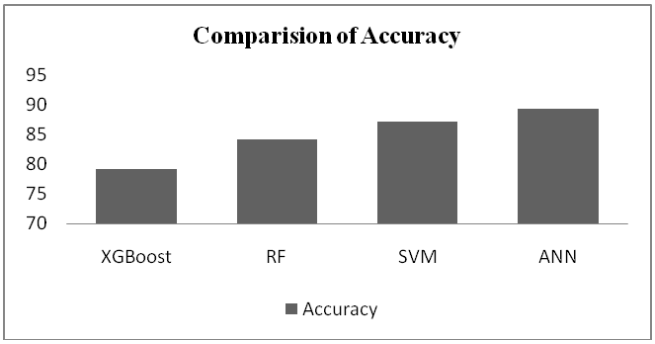

Fig 4 results of accuracy measure

Based on the values obtained from confusion matrix of different methods taken into consideration Recall, Precision, F1 Score and Accuracy are calculated with given equations and shown in Table-3. However, It is identified that ANN is overwhelming with 89.31% other approaches XGBoost (79.12%), Random Forest (84.21%) and support vector machine (87.21%) illustrated in Fig-4.

V.CONCLUSION

Cardiovascular diseases affect the functioning of the heart. There are various approaches including machine learning and soft computing have been applied on existing datasets to get accurate prediction of heart diseases. This experiment provides the deep insight into machine learning techniques for classification of heart diseases. After applying numerous models, and defining log loss as the evaluation metric, we have found the ensemble of SVM, XGBoost, random forest and ANNs to be the best model for achieving the metric laid down. Due to the small size of the dataset given in the competition, Neural Networks could overfit quite easily, and the regularization applied to counter that would bring down the accuracy by quite a bit, resulting in relatively simpler machine learning algorithm like SVM outperforming them. Averaging the probabilities out was the best way to get the least log loss even if the individual models didn't give the best log loss by themselves. As the future work Recurrent Neural Network (RNN) or Long Short Term Memory (LSTM) can be applied as core model while to get optimized parameters one of nature inspired algorithm like Genetic Algorithm, PSO etc.

Table-1 Sample dataset

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	class
63	1	1	145	233	1	2	150	0	2.3	3	0	6	absence
67	1	4	160	286	0	2	108	1	1.5	2	3	3	presence
67	1	4	120	229	0	2	129	1	2.6	2	2	7	presence
37	1	3	130	250	0	0	187	0	3.5	3	0	3	absence
41	0	2	130	204	0	2	172	0	1.4	1	0	3	absence
56	1	2	120	236	0	0	178	0	0.8	1	0	3	absence
62	0	4	140	268	0	2	160	0	3.6	3	2	3	presence
57	0	4	120	354	0	0	163	1	0.6	1	0	3	absence
63	1	4	130	254	0	2	147	0	1.4	2	1	7	presence

Table-2.1 Summary of DataSet

age	sex	cp	trestbps	chol
Min. :29.00	Min. :0.0000	Min. :1.000	Min. : 94.0	Min. :126.0
1st Qu.:48.00	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:120.0	1st Qu.:211.0
Median :56.00	Median :1.0000	Median :3.000	Median :130.0	Median :241.0
Mean :54.44	Mean :0.6799	Mean :3.158	Mean :131.7	Mean :246.7
3rd Qu.:61.00	3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:140.0	3rd Qu.:275.0
Max. :77.00	Max. :1.0000	Max. :4.000	Max. :200.0	Max. :564.0

Table-2.2 Summary of DataSet

fbs	restecg	thalach	exang	oldpeak
Min. :0.0000	Min. :0.0000	Min. : 71.0	Min. :0.0000	Min. :0.00
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:133.5	1st Qu.:0.0000	1st Qu.:0.00
Median :0.0000	Median :1.0000	Median :153.0	Median :0.0000	Median :0.80
Mean :0.1485	Mean :0.9901	Mean :149.6	Mean :0.3267	Mean :1.04
3rd Qu.:0.0000	3rd Qu.:2.0000	3rd Qu.:166.0	3rd Qu.:1.0000	3rd Qu.:1.60
Max. :1.0000	Max. :2.0000	Max. :202.0	Max. :1.0000	Max. :6.20

Table-2.3 Summary of DataSet

slope	ca	thal	class
Min. :1.000	?: 4	?: 2	absence :164
1st Qu.:1.000	0:176	3:166	presence:139
Median :2.000	1: 65	6: 18	
Mean :1.601	2: 38	7:117	
3rd Qu.:2.000	3: 20		
Max. :3.000			

REFERENCES

1. https://www.who.int/news -room/fact -sheets/detail/cardiovascular -diseases -(cvds)
2. K. Polaraju, D. Durga Prasad, “Prediction of Heart Disease using Multiple Linear Regression Model”, International Journal of Engineering Development and Research Development, ISSN:2321 -9939, 2017.)
3. Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36.4 (2009): 7675 -7680.
4. Anbarasi, M., E. Anupriya, and N. C. S. N. Iyengar. "Enhanced prediction of heart disease with feature subset selection using genetic algorithm." International Journal of Engineering Science and Technology 2.10 (2010): 5370 -5376
5. Zhang, Yan, et al. "Studies on application of Support Vector Machine in diagnose of coronary heart disease." 2012 Sixth International Conference on Electromagnetic Field Problems and Applications. IEEE, 2012.
6. Ajam, Noura. "Heart Diseases Diagnoses using Artificial Neural Network." Network and Complex Systems 5.4 (2015): 7 -10
7. Elshazly, Hanaa Ismail, Abeer Mohamed Elkorany, and Aboul Ella Hassanien. "Lymph diseases diagnosis approach based on support vector machines with different kernel functions." 2014 9th International Conference on Computer Engineering & Systems (ICCES). IEEE, 2014
8. Dey, A., Singh, J. and Singh, N., 2016. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis.Analysis,140(2), pp. 27 -31.
9. Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., 2017. Can machine -learning improve cardiovascular risk prediction using routine clinical data?.PloS one,12(4), p.e0174944
10. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.
11. David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database"

AUTHORS PROFILE

Prof. Raghavendra Kumar is working as an assistant professor in department of information technology, KIET Group of institutions Ghaziabad. He has published various articles and research papers in the field of data science, machine learning and deep learning.

Mr. Ashish Mishra is an undergraduate scholar in department of information technology, KIET Group of institutions Ghaziabad. He is passionate to work in the field of machine learning and deep learning.

Mr. Himanshu Rathore is an undergraduate scholar in department of information technology, KIET Group of institutions Ghaziabad. He is passionate to work in the field of machine learning and deep learning.