https://doi.org/ 10.48047/AFJBS.6.7.2024.1632-1641

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Analysing the Performance of Various Machine Learning Techniques in Heart Disease Prediction

Dr.B.Hemalatha,
Department of IT, Dr.N.G.P. Institute of Technology, Coimbatore-641 048

Article History

Volume 6,Issue 7, 2024

Received: 29 Mar 2024

Accepted : 22 May 2024

doi: 10.48047/AF5BS.6.7.2024.1632-1641

Abstract:
One of the main causes of death in the modern days is due to heart related disease. Major problems in clinical data analysis is the accurate prediction of cardiovascular disease. Making judgments and forecasts from the huge amounts of data processed by the healthcare sector has been demonstrated to be aided by machine learning (ML). Early detection of heart failure can be prevented with the accurate and prompt diagnosis of human heart disease, which also increases the prognosis of the patient. Manual methods to diagnose cardiac disease are prone to bias and interexaminer variability. Machine learning algorithms are effective and trustworthy tools for identifying and classifying heart disease patients and healthy individuals. According to various research, only a small portion of heart disease may be predicted using ML approaches. In this research, we suggest a unique approach that aims to increase the optimal prediction rate of cardiovascular illness by utilising machine learning approaches to identify key traits. We introduce SVM- with linear random forest for efficient prediction.

Key terms: Machine learning, diseases prediction, health care, hybrid ML, heart rate detection.

1. Introduction

According to a World Health (WHO) Organization estimate, 17.91 million fatalities across world in 2019 were attributable to cardiovascular diseases (CVD) [1], accounting for 32.01% of deaths worldwide [2]. The annual mortality rate for CVD was higher than 17.70 million. According to the Australia Institution of Healthcare and Welfare (AIHW) of cardiovascular disease

(CVD) constituted for 42.01% of all fatalities in Australia during 2018 and was the high cause of death [3]. Because current approaches for predicting heart disorders are not accurate or computationally efficient for early detection, researchers are working to develop a system that will enable the prompt identification of heart diseases [4]. The identification and management of cardiac disease are difficult when cutting edge technologies and health professionals are not available [5].

It is hard to predict heart problems because of the many complex factors that contribute to it, such as diabetes,excessive cholesterol, high pressure, an irregular pulse rhythm, and many more. The degree of heart disease in individuals has been assessed using a variety of data mining and neural network techniques. The severity of the ailment is classified using a variety of methods, such as Decision Trees (DT), K-Nearest Neighbor ,Genetic Algorithms (GA), and Naive Bayes (NB) [11], [13]. Heart disease requires appropriate treatment due to its complex nature. Failing to take action so could cause early death or harm to the heart. Through the use of data mining and a medical research perspective, the many forms of metabolic disorders are identified.

Machine learning (ML) in the prediction of heart diseases leverages advanced algorithms to analyze a wide array of data—from patient demographics and clinical records to lifestyle choices and genetic markers. By applying techniques such as logistical regression, decision tree, and neural networks, ML models can discern complex patterns and risk factors associated with cardiovascular conditions. This not only aids in early diagnosis and personalized treatment plans but also enhances the overall effectiveness of healthcare interventions. As this technology evolves, it holds the promise of transforming cardiac care, ensuring better patient outcomes through predictive analytics and data-driven insights

We also discuss the use of Computer Aided Decision Support model (CADSM) in research and medicine. Previous research has demonstrated that the application of data mining techniques in the healthcare sector can forecast disease more accurately and in a shorter amount of time [16]. We suggest using the genetic techniques to predict cardiac disease. Effective association rules for crossover, tournaments selection, and mutate—which yields the new suggested fitness function—are inferred using the GA in this method. For experimental , The Cleveland dataset is used by collecting from the UCI repository in machine learning. Later on, results are analysed comparison to few of the well-known supervised techniques [5].

Important risk factors that current models do not fully account for. Our research will concentrate on combining the conventional clinical signs with a wider range of variables, such as environmental and socioeconomic factors. It is expected that this integration will produce a more complete model than the existing models, one that better reflects the subtleties driving heart disease risk. In addition, we want to utilize sophisticated machine modeling methods like deep learning and ensemble learning to enhance the precision and dependability of our forecasts. Our goal is to close these gaps so that our study can improve ML models' predictive power while also lessening the biases and inconsistencies that come with manual diagnostic procedures. In the end, this might result in actions that are more accurate and timelier, greatly enhancing patient outcomes in cardiovascular care.

For the purpose of predicting heart disease, we developed a machine learning classifier in this study that combines several machine learning techniques, such as logistical regression (LR),multinomial Na¨ıve Bayes (MNB), logistic discriminant analyser (LDA), XGBoost (XGB), support vector machine (SVM), decision trees (CART), extra trees classifier (ET) and random forest (RF),. The hyperparameters and standrad data are set using the GridSearch method in order to determine the optimal hyperparameter to the most efficient machine learning classifier. Aside from that, the machine learning classifier's performance is assessed using a range of performance evaluation metrics, including F-measures, sensitivity, recall, accuracy, and precision. The HD Cleveland dataset has been used to test the suggested approach. Additionally, the suggested classifiers accuracy has been contrasted with current state-of-the-art techniques in the comparison.

2. Literature survey

In the fields where this study is directly relevant, there is a wealth of related work. In order to provide predictions with the greatest degree of precision in the medical industry, ANN was introduced [6]. The backpropagation multilayer perception predicts heart disease (MLP) of ANN. When the acquired findings are contrasted with those of previous models in the same field, they show improvement [10]. NN, Support Vector machines, DT,, and Naive Bayes are utilized to find patterns in the patient data from the UCI laboratory related to heart disease. With these algorithms, the accuracy and performance of the outcomes are compared. In terms of F-measure, the suggested hybrid approach competes with the other current techniques, yielding values of 86.8% [7]. Previously, a essential data generated by the medical field was not utilized efficiently. The innovative methods offered here decreases costs and enhances prediction of heart diseases in simple and effective manner. The more research models taken into consideration in this work for the deep learning (DL) and machine learning (ML)- based prediction and classifier of heart strokes are quite accurate in demonstrating the effectiveness of these approaches [15].

A thorough heart disease presence can be predict based on analysis of mostly used machine learning classifiers which was employed in a different study carried out in [16] Just 14 features—out of the 303 records in the Cleveland (UCI) datasets—are used for training data and testing data. Following the completion of data preparation, a dataset with 296 records was produced. The accuracy of the SVM classifier findings was greater, at 90.00%. In a study to predict cardiac disease, [12] used hybrid models of data mining classifier. The UCI repository of machine learning provides the datasets, which have 76 attributes and 303 records. training of the model and testing were done on 14 attributes.

The [9] description of the current and prospective state of AI-enhanced electrocardiograms (ECGs) in at-risk communities' heart disease diagnosis, summary of its implications for patients' healthcare decisions, and evaluation of its possible downsides. A novel health information system was presented in [10] to prescribe exercise to individuals with heart disease. Their preliminary research indicates that physicians are unsure of how to create an exercise prescription. Mobile Information Systems 39071, 2022, 1, retrieved [13/06/2024] in https://onlinelibrary.wiley.com/doi/10.1155/2022/1410169, Wiley Online Library.

3. Proposed Methodology

In this work, we differentiate heart diseases data from Cleveland UCI repository using a R studio rattle. It offers a easy operation of visual monitoring of the dataset, the workspace, and the predictive analytics process. The machine learning method begins with pre-processing the data, then moves on to feature selection using DT entropy, modeling classification evaluation performance, and better accuracy result. The process of selecting features and modelling never ends for different sets of attributes.

3.1 Data collection

The most well-known dataset that the researchers have utilized is the Cleveland dataset on heart disease, which can be accessed through the University of California, Irvine (UCI) online machine learning repository. There are totally 303 records, and samples of 6 lack values. The original version of the +e data contained 76 features; however, only 13 of these are likely to be mentioned in published work, with the remaining feature detailing the disease's presence. +e Z-Alizadeh Sani dataset is another well-liked dataset that researchers used in the prediction procedure. It contains 303 data about the patients with 55 input parameters and a each patient with class label variable. The researchers also employed Hungarian, , Kaggle Framingham, Long Beach VA and StatLog Heart datasets in their prediction process[13-15]. Table 1: Dataset details and samples.

Datasets	**Samples**	**Risk factors**	**reference**
UCI	303	13	[13]
Z-Alizadeh Sani	303	55	[14]
StatLog	270	13	[15]

.

3.2 Data standardization

We improved and standardized the data sets that were gathered. These datasets had inaccurate values and were not collected in a controlled setting. For this reason, preparing data is a hard stage in the data analyzing using machine learning. Normalization of dataset refers to the process of a dataset's risk factors having distinct values. For instance, the temperature can be measured in multiple ways using Celsius and Fahrenheit. Scaling the risk variables and allocating numbers that illustrate the variation in standard deviation from the mean value constitute the process of standardizing the data. To enhance the machine learning classifier performance with a mean (μ) of 0 and a standard deviation (σ) of 1, it rescales the risk factor value. (1) provides the standards in mathematical form.

$$Data\ standardization\ X = \frac{X - Mean\ of\ X}{Standard\ deviation\ X} \tag{1}$$

3.3 Parameter processing with SVM-Linear Random forest

To achieve high precision, hyperparameter tuning is used to determine the ideal value for the hyperparameters. We employed the GridSearchCV technique for this. To improve the effectiveness of machine learning classifiers, we modify their hyper parameter values before to using them. The fit method of the Scikit-learn GridSearchCV class offers a grid of tweaking classification methods. It makes it possible to train any machine learning algorithm in a single, reliable environment and to modify the corresponding hyperparameters. Once the appropriate values for the hyperparameters have been found, the full training dataset is used to create an accurate model.

As was already noted, a number of (ML) techniques—GLM, NB, LR, RF, DT, GBT, and SVM—are employed in this experiment. All 13 attributes were used in the experiment, which was conducted again using every ML technique.

Although it's more often employed for classification problems, SVM seems strong, adaptable supervised learning model utilized for regression as well. The idea behind SVM is to know the ideal hyperplane for classifying distinct groups of data. The hyperplane that maximizes the margin between the closest points in each class— a.k.a. support vectors—is the ideal one. The algorithm seeks to maximize this margin, which is thought of as a separating zone. The decision line in the feature space that divides several classes is called a hyperplane. This hyperplane can be seen as a line in two dimensions. The support vectors are the data points that are closer to the hyperplane. These points occupy main role in determining the hyperplane's position since they directly affect its orientation and shape. On the nearest class points, this is the distance between the two lines. A higher margin suggests that the classifier's generalization error is smaller.

When working with non-linear boundaries in SVM, the kernel technique is a crucial idea. It entails converting data into a higher dimension where class separation can be achieved with just a linear separator. Typical kernels consist of: Ideal for data that can be divided linearly, the linear kernel requires no change.

Model Trees, in which linear regression models are fitted at the leaves rather than utilizing the mean or mode of the target variable, could be one method if by "Linear Random Forest" you imply integrating linear models within a Random Forest-like framework. This approach is more frequently linked to machine learning methods such as M5 trees, in which the data space is partitioned using decision trees and linear models are fitted to each partition.

Partitioning: A decision tree structure is used to divide the data, but a linear regression model is fitted rather than a prediction of the outcomes at each leaf.

Data splitting: Splits are made using standard criteria, such as variance reduction (for regression jobs).

4. Results and Discussion

After gathering the dataset from repository of machine learning, we cleaned, standardized, and improved it. Following normalizing data, we used machine learning classifiers and hyperparameter optimization. 10 fold cross validation is done for training and testing all of the classifiers. Additionally, classifier accuracy is examined both before and after datasets standardization. Plotting the chosen classifiers accuracy is done for evaluation purposes. The classifiers' accuracy before and after the normalization of the data is displayed in Figure 2. Figure 2 shows that on the standardised dataset, the majority of machine learning approaches (RF, LDA, CART,LR, ET, AB,and XGB) increased accuracy. Others like MNB and SVM methods lowered accuracy. On the standard dataset, several classifiers, including ET, CART,and AB, shown notable gains in accuracy.

accuracy of the classifier in prediction is calculated using,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The projected positive cases that are actual or true positives are measured by precision. In terms of math, it is provided by

$$Precision = \frac{TP}{TP + FP}$$

The overall number of actual positive cases as impacted by the false negative instances in total is analyzed by recall.

$$Recall = \frac{TP}{TP + FN}$$

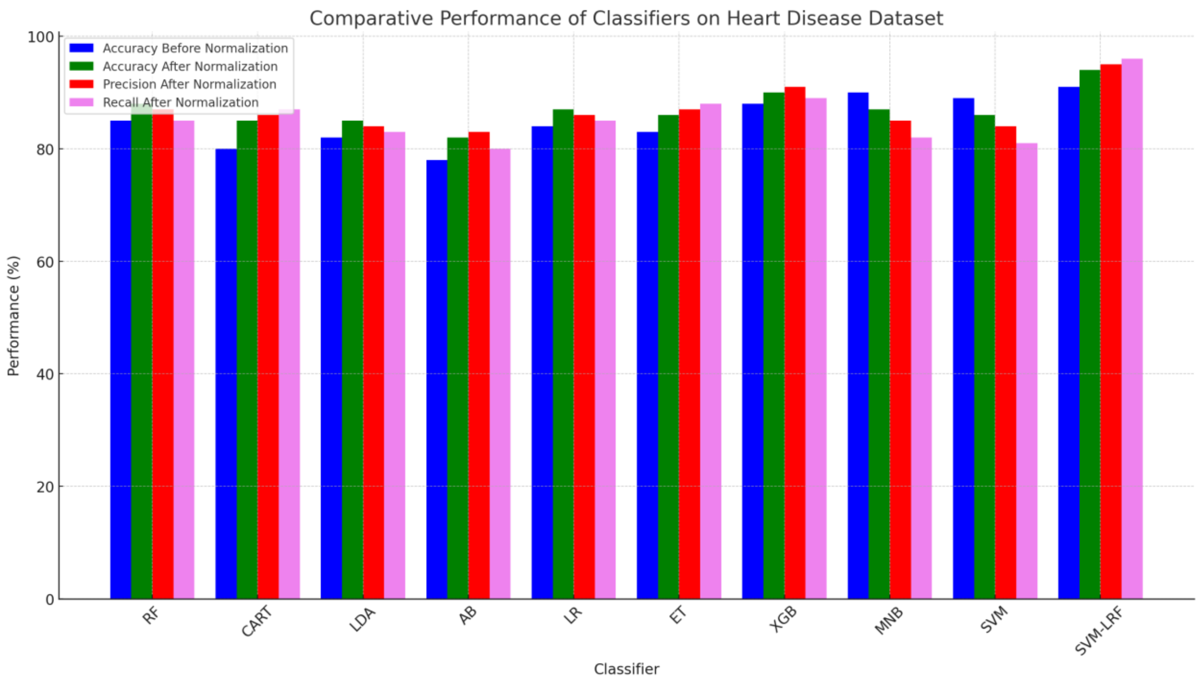

Figure 1: comparison of Various ML with proposed SVM-LRF

4. Conclusion

Long-term life preservation and the early identification of anomalies in cardiac conditions can be facilitated by knowing the raw healthcare data processing related to cardiac information. In this study, raw data was executed using machine learning to produce a fresh and innovative diagnosis of heart disease. In the automation diagnosis system, heart disease prediction is challenging and also important life saving aspect. However, the aim of decreasing death rate is possible by detecting illness in early and preventative measures are implemented as soon as practical. It would be ideal for this study to be expanded upon research on real time dataset rather than theoretical approaches and simulations. The suggested hybrid HRFLM technique combines Random Forest features with linear method.

References:

[1] A. K. Dwivedi, S. A. Imtiaz, and E. R. Villegas, "Algorithms for automatic analysis and classification of heart sounds – a systematic review," IEEE Access, vol. 7, 2019.

[2] A Coronary, "Heart disease," Available from: https://www. aihw.gov.au/reports/australias-health/coronaryheart-disease, 2020.

[3] L. A. Allen, L. W. Stevenson, K. L. Grady et al., "Decision making in advanced heart failure: a scientific statement from the American heart association," Circulation, vol. 125, no. 15,pp. 1928–1952, 2012.

[4] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," Journal of Intelligent Learning Systems and Applications, vol. 5, no. 3, Article ID 35396, 2013.

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

[6] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig,"Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," Journal of 'e Royal Society Interface, vol. 8, no. 59, pp. 842–855, 2011.

[7] S. I. Ansarullah and P. Kumar, "A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method," International Journal of Recent Technology and Engineering, vol. 7, no. 6S,

pp. 1009–1015, 2019.

[8] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, "Fuzzy logic based decision support system for component security evaluation," 'e International Arab Journal of Information Technology, vol. 15, no. 2, pp. 224–231, 2018.

[9] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," Journal of Intelligent

Learning Systems and Applications, vol. 5, no. 3, pp. 176–183,2013

[10] S. Mohan, C. +irumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019.

[11] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi, and A. S. Sulthana, "Analysis and prediction of cardio vascular disease using machine learning classifiers," in Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 15–21, IEEE, Coimbatore, India, 2020 March.

[12] Cleveland Clinic Foundation, "Heart disease data set,"Available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[13] Statlog, “Heart disease data set,” Available at: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29.

[14] Z. Alizadeh Sani, “Heart disease data set,” Available at: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani.

[15] 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 Proceedings of the IEEE Symposium Computer Communication (ISCC), pp. 204–207, Heraklion, Greece, July 2017.