Acta Scientific COMPUTER SCIENCES

Volume 6 Issue 7 July 2024

Research Article

Refined Approach for Predicting Heart Disease Through Machine Learning and Feature Engineering Techniques

Sunil Bhutada, K Usharani, K Mahesh*, L Manish Reddy, N Sravan Kumar and D Vaishnavi

Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India

***Corresponding Author:** K Mahesh, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India.

Received: May 17, 2024

Published: June 04, 2024

Abstract

The importance of early diagnosis is highlighted by heart disease, a chronic illness that affects millions of people worldwide. To isolate and improve upon the most important characteristics, a novel feature engineering approach is presented that makes use of Principal Component analysis. The study's goal is to develop a user interface and use machine learning (ML) to quickly anticipate the health condition of heart disease and start the necessary steps. A Stacking Classifier, an ensemble approach, is used in the work to integrate the predictions of three different models: Random Forest (RF), Multilayer Perceptron (MLP), and Light GBM. By combining the best features of many models in a complementary fashion, this method achieves a remarkable 100% accuracy rate in its final forecast, making it both resilient and precise. Model construction made use of the characteristics chosen according to Principal Component Heart Failure (PCHF), and front-end deployment of the Stacking Classifier was trained, have improved the accessibility and usefulness of our machine learning-based heart disease prediction system by integrating the Flask framework with user authentication. The patient will enter the details at a user interface and determine whether the user has heart disease or not. In this study comparative analysis of classification algorithms along with ROC curve of the different classification techniques. This offers a safe and effective platform for user testing.

Keywords: ML; Cardiac; Arrest; Cross-Validations; Feature; Engineering; Algorithms; Patient; Prevalence; Accuracy

Introduction

In heart failure, the heart's ability to pump blood is inadequate to fulfill the body's demands [1]. The prevalence and severity of cardiovascular illnesses have recently grown into a major issue in public health across the globe. Heart failure is a major health concern that impacts millions of people around the globe. About 26 million people are affected by heart failure problems, according to a recent survey [2]. There are two main types of causes of heart failure. Anatomical issues, such as a history of heart attacks, are the first to be considered. Second, conditions that affect the heart's ability to pump blood, such as hypertension. Breathlessness, lethargy, and edema in the lower extremities are signs of heart failure. Heart failure treatment methods may include medication, behavioral modifications, and, in rare instances, surgical intervention. Detecting and treating heart failure early may increase survival rates and enhance quality of life, according to research [3]. To better manage heart failure and enhance patient health, the present work aims to construct an ML model.

The healthcare business and medical diagnosis rely heavily on ML [4]. In medicine, ML finds several uses, such as in the development of new drugs, the interpretation of medical images, the forecasting of epidemics, and the detection of heart failure. The use of ML methods allows for the predictive analysis and pattern recognition of massive amounts of medical data. When compared to traditional medical practices, ML offers several benefits, such as better diagnosis with less expense and less time spent on testing.

To choose the most important features to improve performance, a new PCHF feature engineering method is suggested. The ML algorithms that use the suggested PCHF methodology are based on eight aspects of the dataset that have high relevance values. To improve upon previous methods, developed a new set of features to enhance the suggested PCHF mechanism and get the best possible accuracy results. To forecast cases of heart failure, the comparison makes use of nine sophisticated ML algorithms. To get a high-performance accuracy score, the hyperparameters of each ML

algorithm are tuned to find the best-fit parameters. Using the k-fold cross-validation approach, have validated the performance of applicable ML models.

The states mentioned in the prior research all agree that heart disease is the deadliest human illness. Healthcare systems around the globe are facing a major challenge and danger from the rising prevalence of deadly cardiovascular illnesses [5]. Most people afflicted by this serious illness are children [6]. Classification models have been used in healthcare before, and [7] explains what such models are and why they are useful. Multiple research groups have shown promising results when using data mining techniques in healthcare settings, according to the report. Experts in the field used two tools, KA, and MATLAB, to evaluate several functional classifiers. Algorithms such as decision trees, logistic regression, support vector machines, and others achieved relatively low accuracy levels ranging from 52% to 67.7% [8].

The accuracy was enhanced from 87.27% to 93.13% by the previous study [9], which is excellent but not optimum, as shown in Table 1. Studies conducted in the past have used techniques including support vector machines, RFs, decision trees, logistic regression, and naïve Bayes classifiers to identify patients with heart failure. The decision tree produced a decent identification of heart failure in a particular dataset with an accuracy of 93.19% after comparing the findings.

An ensemble model for the identification of cardiac disease was developed using data collected from Cleveland in the research [9]. An accuracy of 85.71% was achieved by the ensemble models that re constructed using the classifiers RF, gradient boosting, and extreme gradient boosting. The suggested research employed the Cleveland data to enhance the accuracy of heart disease prediction using a feature selection approach, leading to an 86.60% success rate. Lastly, there are large gaps in the literature that point to subpar performance accuracy, as seen in earlier investigations. As a result, assess the performance analysis of the prior research in this section. Results that summarize the efficacy of all previously used models form the basis of this section on related work. Various models continue to produce varying prediction scores, as shown in earlier research. Data selection may be improved, leading to more accurate predictions, via dimensionality reduction and feature engineering [10].

The accuracy score of our suggested investigation is higher than that of our earlier research. Accurate diagnosis and evaluation of heart failure is critical for effective therapy. To do this, used state-of-the-art ML methods in our research.

Literature Review

More than 26 million people throughout the globe are living with chronic heart failure, making it a true epidemic. It causes about a million hospitalizations yearly in North America and Europe and is a leading cause of mortality among individuals with cardiovascular illnesses. Patients' quality of life may be greatly improved by using methods for chronic heart failure detection to take preventative measures, improve early diagnosis, and avoid hospitalizations or potentially life-threatening circumstances. Our machine-learning approach for detecting chronic heart failure from cardiac sounds is presented in this publication. The steps in the process include using filters, segmentation, feature extraction, and ML [11]. Using information from 122 participants, the strategy was evaluated using a leave-one-subject-out design. The approach outperformed a majority classifier by 15% with an accuracy of 96%. Specifically, it remembers 87% of the people with chronic heart failure and recognizes them with 87% accuracy. The study's findings validate the feasibility of using sophisticated ML to diagnose chronic heart failure using real-life sounds captured with an inconspicuous digital stethoscope.

A growing number of individuals are being affected by heart failure (HF), which is a worldwide epidemic that has already affected 26 million people. The costs associated with heart failure are high and are projected to rise substantially as the population ages. There has been great progress in treatment and prevention, but quality of life is low and mortality and morbidity are still high. Geographic differences in reported rates of prevalence, incidence, mortality, and morbidity are attributable to differences in the etiologies and clinical features of HF patients [12]. Here provide statistics on the incidence, prevalence, mortality, and morbidity of HF on a global scale, with an emphasis on the epidemiology of the disease.

The improved performance of ML and deep learning (DL) methods in several healthcare applications, such as CADx (computer-aided diagnosis) with multi-dimensional medical images and cardiac arrest prediction from one-dimensional heart signals, has led to their broad adoption in recent years. Recent findings have demonstrated that ML/DL are susceptible to adversarial attacks, adding fuel to the fire of lingering concerns about the robustness of ML/DL in healthcare settings, where the myriad security and privacy issues make it traditionally considered quite challenging. The inherent problems provide an overview of several healthcare application areas that use these approaches from a security and privacy perspective. also provide some possible solutions to the problem of how to make ML for healthcare applications safe and private. Lastly, provides light on the present difficulties of the field and suggests interesting avenues for further study.

The prediction of cardiac illness by using eleven ML classifiers to discover critical characteristics. The prediction model was introduced using standard classification algorithms and a variety of feature combinations. Our heart disease prediction model reached 95% accuracy using gradient-boosted trees and MLP. The RF outperforms the other methods in predicting the occurrence of cardiac problems, with a 96% success rate.

People nowadays are so busy with work and other responsibilities that they forget about taking care of themselves. Every day, more and more individuals become ill because of their careless attitude toward their health and the frenetic pace of modern life. In addition, many are afflicted with ailments such as cardiovascular disease. According to statistics provided by the World Health Organization (WHO), heart disease is responsible for the deaths of over 31% of the world's population. As a result, the ability to forecast the occurrence of heart disease takes on more significance in the medical industry. The problem is that hospitals and the medical industry get massive amounts of data, which may be difficult to assess at times. The medical community may greatly benefit from the use of ML methods in this area of data processing and prediction. Therefore, have covered the causes of heart disease and the variables that put people at risk, as the methods of ML [13]. have utilized these ML methods to forecast the occurrence of heart disease and have compared the various ML algorithms that are used in this experiment. The purpose of this study is to forecast the occurrence of cardiovascular illness utilizing an ML approach and its analysis.

Methodology

To better diagnose heart disease at an early stage, ML technologies are used. use and compare nine different ML algorithms: logistic regression, RF, support vector machine, decision tree, extreme gradient boosting, naive Bayes, KNNs, MLP, and gradient boosting. A novel feature engineering approach is presented that emphasizes the identification of crucial features using Principal Component Analysis (PCA), to improve accuracy. A Stacking Classifier, an ensemble approach, is also used; it integrates the forecasts of the RF, MLP [14], and LightGBM models. By combining the best features of many models in a complementary fashion, this method achieves a remarkable 100% accuracy rate in its final forecast, making it both resilient and precise. Model construction made use of the characteristics chosen according to PCHF, and front-end deployment of the Stacking Classifier was trained. have improved the accessibility and usefulness of our machine learning-based heart disease prediction system by integrating the Flask framework with user authentication [15].

The Kaggle repository provided us with the heart failure dataset that was used in this investigation. There are 1025 records in the collection, and they pertain to both healthy and heart failure patients. To format the dataset, the data preparation procedures are used. If you want to learn more about the factors that lead to heart failure and the patterns in the data, you should use exploratory data analysis. Feature engineers use the suggested PCHF method to pick out the most important features. Subsequently, the dataset is partitioned into two parts: train and test. apply the nine state-of-the-art ML methods to the dataset sections. fine-tune the ML models using hyperparameters. The goal of the newly suggested approach is to provide very accurate predictions of cardiac failure.

ML algorithms are trained and tested using the heart disease dataset [16], which includes detailed clinical and patient data, such as demographics, medical history, and physiological measures, to accurately predict the occurrence of heart disease.

Table 1: Sample data of dataset.

S no	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Old peak	Slope	Ca	Thal	Target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	2	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	1	2	0

Data processing

Processing data entails making sense of raw data for companies. Collecting, organizing, cleaning, validating, analyzing, and transforming data into understandable representations like graphs or papers are all part of data processing. There are three main ways that data may be processed: mechanically, electronically, or by hand. Improving the usefulness of data and making decisions easier are the goals. Companies may then use this information to make better strategic choices and enhance their operations. Software development and other forms of automated data processing are crucial here. Quality management and decision-making may benefit from its ability to transform massive data sets, particularly big data, into actionable insights.

Feature selection

It refers to the process of identifying which characteristics are most relevant, consistent, and free of duplication before building a model. With the proliferation of datasets comes the need to systematically reduce their sizes. The primary objective of feature selection is to decrease the computational cost of modeling while simultaneously improving the performance of a predictive model.

An essential part of feature engineering is feature selection, which entails picking out the most relevant characteristics to feed into ML algorithms. By removing superfluous or unimportant features and keeping just the most important ones, feature selection approaches help to decrease the number of input variables needed by the ML model. Rather than relying on the ML model to prioritize features, it is recommended to undertake feature selection beforehand.

Classification and predictive analytics often use logistic regression models. Using a collection of independent variables, logistic regression may determine the likelihood of an event happening, such as voting or not voting, according to some publications [17].

DT

Decision trees are used for both regression and classification applications; they are non-parametric supervised learning algorithms. Its structure is organized like a tree, with a root node, branches, internal nodes, and leaf nodes.

The RF method

It is a popular ML technique that uses a series of decision trees to get a result. It manages both classification and regression issues, and its adaptability and user-friendliness have contributed to its widespread usage [18].

SVM

This is an effective supervised technique that excels at handling complicated datasets on smaller ones. While SVMs are versatile, they often shine when used for classification issues rather than regression [19].

A non-parametric supervised learning classifier, the k-nearest neighbors (KNN) method analyzes the degree of closeness between two points in a dataset to conclude how those points should be grouped.

MLPs are a kind of current feedforward artificial neural networks that are characterized by their ability to differentiate input that is not linearly separable. These networks are composed of fully connected neurons with a nonlinear activation function and are structured in at least three layers. The term is misleading as current networks utilize nonlinear activation functions, although the original perceptron relied on a Heaviside step function.

ML model training has never been easier than with XG Boost, an improved distributed gradient boosting toolkit. As an ensemble learning technique, it takes the predictions of several models and merges them into a single, more robust prediction.

One well-known boosting technique in ML for classification and regression applications is Gradient Boosting. One kind of ensemble learning is boosting, which involves sequentially training the model to improve upon earlier iterations. It turns several poor students into a few excellent ones [20].

Ensemble learning is the stacking classifier, which takes several classification models and merges them into a single "super" model. Since the combined model may benefit from each model's capabilities, this frequently results in better performance.

Experimental results

Accuracy: A test's accuracy is defined by how it distinguishes between healthy and sick samples. can determine a test's accuracy by calculating the percentage of revied instances with true positives and true negatives.

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

Precision

The accuracy rate, or precision, is the percentage of true positives relative to the total number of occurrences or samples. Consequently, the following is the formula for determining the accuracy: Preciseness is TP divided by (TP plus FP), which is the sum of true positives and false positives.

Recall

The capacity of a model to detect all significant occurrences of a given class is measured by the recall, a statistic in machine learning. The completeness of a model in capturing instances of a particular class is shown by the ratio of properly predicted positive observations to the total actual positives.

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

F1-Score

One way to evaluate a model's performance in ML is via its F1 score. This method integrates a model's recall and accuracy scores. A model's accuracy may be measured by counting the number of times it correctly predicted something throughout the whole dataset.

ML MODEL	Accuracy	F1 SCORE	Recall	Precisions
Logistic regression	0.75	0.816	0.724	0.767
Decision tree	1	1	1	1
Random forest	1	1	1	1
SVM	0.639	0.66	0.636	0.648
KNN	0.902	0.193	0.895	0.904
MLP	0.59	0.184	1	0.311
Naïve bayes	0.751	0.777	0.741	0.758
Xg boosting	0.912	0.913	0.913	0.913
Gradient boosting	0.922	9.22	0.922	0.922
Staking classifier	1	1	1	1

Table 2: Performance analysis of different classifications.

ROC curve

The receiver operating characteristic (ROC) curve is a graphical tool used to assess the performance of quantitative diagnosis methods for categorizing. The tool differentiates between typical and atypical outcomes, displays the sensitivity and specificity for each threshold, and is not influenced by the occurrence. Nevertheless, does not have a clear presentation of the distinction between typical and abnormal cut-off values, as well as the number of samples. The curve has a fragmented appearance when the sample size is less, but becomes smoother as the number of samples increases.

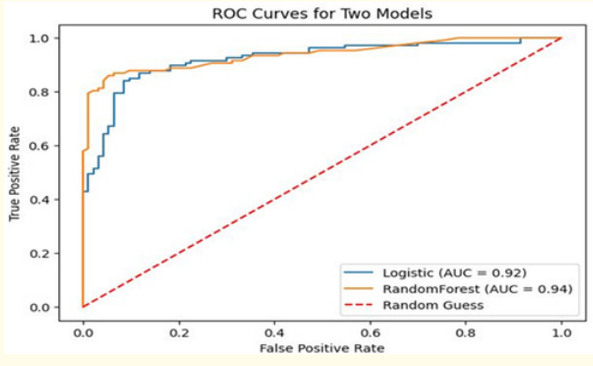

Figure 1: ROC for LR and RF classification techniques.

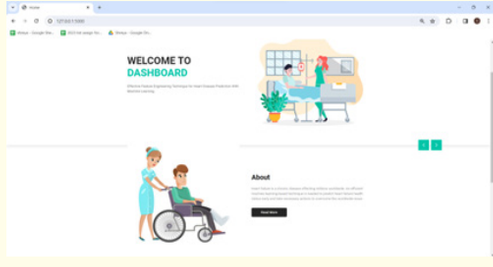

Figure 2: The user interface of the proposed method.

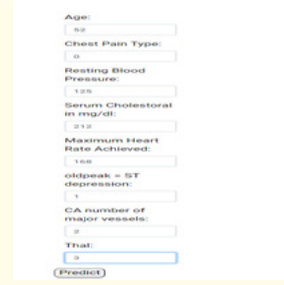

Figure 3: A user interface to enter the patient details for predicting.

You have no Heart Disease, based on the input provide!

Figure 4: Predicting results of uploading patient details.

Conclusion

A strategy for predicting cardiac arrest using ML algorithms. The models that are put into use are built using a dataset that is based on 1025 patient records. An innovative method for PCHF feature engineering is suggested, which prioritizes eight key characteristics to improve performance. When comparing the ML approaches that were reused, many came up: decision trees, logistic regression, RFs, SVMs, extreme gradient boosting, naive base, KNN, MLP, and gradient boosting. With just 0.005 calculations performed during runtime, the suggested DT technique achieved 100% accuracy. Each learning model has its performance validated using the cross-validation approach that relies on 10-fold data. have developed a generalizable approach for identifying heart failure that outperformed the state-of-the-art research.

Future Enhancement

Our suggested methodologies have the potential to set a benchmark for performance in the prediction of cardiac disease, which may guide future studies in this area. To improve the efficiency of categorization models, future research may investigate ways to better manage features. In addition, our approach can be used in numerous medical fields to improve the use of ML algorithms for illness prediction and identification.

Bibliography

1. M Gjoreski., *et al*. "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers". in Proc. Int. Conf. Intell. Environments (IE), Aug. (2017): 14-19.

2. G Savarese and L H Lund. "Global public health burden of heart failure". *Cardiac Failure Review* 3.1(2017): 7.

3. EJ Benjamin., *et al*. "Heart disease and stroke statistics—2019 update: A report from the American Heart Association". *Circulation* 139.10 (2019): e56-e528.

4. A Qayyum., *et al*. "Secure and robust machine learning for healthcare: A survey". IEEE Rev. Biomed. Eng. 14 (2021): 156-180.

5. C Trevisan., *et al*. "Gender differences in brain-heart connection". *Brain and Heart Dynamics* (2020): 937-951.

6. DC Yadav and S Pal. "Prediction of heart disease using feature selection and random forest ensemble method". *International Journal for Pharmaceutical Research Scholars* 12.4 (2020): 56-66.

7. D Tomar and S Agarwal. "A survey on data mining approaches for healthcare". *International Journal of Bio-Science and Bio-Technology* 5.5(2013): 241-266.

8. S Ekiz and P Erdogmus. "Comparative study of heart disease classification". in Proc. Electr. Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT) (2017): 1-4.

9. BA Tama., *et al*. "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble". *BioMed Research International* (2020): 1-10.

10. FS Alotaibi. "Implementation of a machine learning model to predict heart failure disease". *International Journal of Advanced Computer Science and Applications* 10.6 (2019): 1-8.

11. V Ramalingam., *et al*. "Heart disease prediction using machine learning techniques: A survey". *International Journal of Engineering and Technology* 7.2 (2018): 684-687.

12. D K Plati., *et al*. "A machine learning approach for chronic heart failure diagnosis". *Diagnostics* 11.10 (2021): 1863.

13. A Saboor., *et al*. "A method for improving prediction of human heart disease using machine learning algorithms". *Mobile Information Systems* (2022): 1-9.

14. R Katarya and S K Meena. "Machine learning techniques for heart disease prediction: A comparative study and analysis". *Health and Technology* 11.1 (2021): 87-97.

15. R Pahuja and A Kumar. "Sound-spectrogram based automatic bird species recognition using MLP classifier". *Applied Acoustics* 180 (2021): Art. no. 108077.

16. B Olimov., *et al*. "Right initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model". *Concurrency and Computation: Practice and Experience* 33.22 (2021): e6143.

17. K Shah., *et al*. "A comparative analysis of logistic regression, random forest and KNN models for the text classification". *Augmented Human Research* 5.1 (2020): 1-16.

18. V Kakulapati., *et al*. "Multimodal detection of COVID-19 Fake News and Public Behavior Analysis- Machine Learning Perspective" is in the book "Springer Innovation in Communication and Computing series "Intelligent Healthcare". (2021).

19. V Kakulapati., *et al*. "A Novel Multimodal Risk Disease Prediction of Covid-19 by Using Hierarchical LSTM Methods", "Taylor and Francis book" Data Science and Data Analytics: Opportunities and Challenges".

20. A Raza., *et al*. "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction". *PLoS ONE* 17.11 (2022): Art. no. e0276525.