Computer Assisted Methods in Engineering and Science
31(4): 419–429, 2024, doi: 10.24423/cames.2024.602

This article belongs to the *Special Issue on Scientific Computing and Learning Analytics for Smart Healthcare Systems* edited by Dr. C. Chakraborty, Dr. S. Barbosa and Dr. L. Garg

Machine Learning and Feature Selection-Enabled Optimized Technique for Heart Disease Classification and Prediction

P. NANCY[1], Prasad Raghunath MUTKULE[2],
Kalpana Sunil THAKRE[3], Ajay S. LADKAT[4],
S.B.G. Tilak BABU[5], Sunil L. BANGARE[6], Mohd NAVED[7]*

[1] *Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattangulathur Campus, Chennai, India; e-mail: nancyp@srmist.edu.in*

[2] *Department of Information Technology, Sanjivani College of Engineering, Kopargaon, Maharashtra, India; e-mail: mutkuleprasadit@sanjivani.org.in*

[3] *Department of Computer Engineering, MMCOE, Savitribai Phule Pune University, Pune, India; e-mail: kalpanathakre@mmcoe.edu.in*

[4] *Department of Instrumentation Engineering, Vishwakarma Institute of Technology, Pune, India; e-mail: ajayladkat123@gmail.com*

[5] *Aditya Engineering College, Surampalem, India; e-mail: thilaksayila@gmail.com*

[6] *Department of Information Technology, Sinhgad Academy of Engineering, Savitribai Phule Pune University, Pune, India; e-mail: sunil.bangare@gmail.com*

[7] *Jaipuria Institute of Management, Noida, India*

* *Corresponding Author e-mail: mohdnaved@gmail.com*

When dealing with a group of patients seeking treatment for heart-related diseases, doctors who specialize in the diagnosis and treatment of heart-related disorders have a difficult but critical task. It comes as no surprise that cardiovascular disease is a leading source of morbidity and death in contemporary society. An expert system with clear categorization that may assist medical professionals in identifying heart disease condition based on the clinical data of a patient is often required by physicians. The aim of this work is to provide a method for the prediction and classification of cardiac disease based on machine learning and feature selection. The correlation-based feature selection (CFS) method is applied to the input data set in order to extract relevant features for analysis. The support vector machine with radial basis function (SVM RBF) and random forest algorithms are used here for data classification. Cleveland heart disease dataset is used in the experiment work. This dataset has 303 instances and 14 attributes. The accuracy, specificity and sensitivity of SVM RBF are higher than those of the random forest algorithm.

Keywords: machine learning, heart disease prediction, accuracy, SVM RBF, CFS feature selection.

1. Introduction

The heart is a very important organ that is a key component of the cardiovascular system in the human body, and cannot be neglected or undervalued. The heart functions as a pump that distributes blood throughout the body. If blood circulation in the body is inadequate, the brain and other essential organs suffer, and in the event of heart failure, death can occur within minutes. Doctors who specialize in the diagnosis and treatment of heart-related ailments have a challenging but critical role when dealing with patients seeking treatment for such conditions. It is no surprise that cardiovascular illness is a primary cause of morbidity and mortality in modern society, given the pivotal role of the heart in sustaining human life [1, 2].

Cardiac disease may be treated if it is detected early enough to reduce the risk of death. In the past, physicians employed a range of medical tests, including electrocardiograms (ECGs), cardiac magnetic resonance imaging (MRI), and stress testing to detect heart disease. However, manual detection, in the majority of situations, necessitates consultation with a specialized physician [3, 4].

Patients can only make educated guesses about their own health if their heart communicates symptoms while they go about their daily activities as usual. A variety of computer-aided expert systems have been developed to assist clinicians in better anticipating the onset of cardiovascular disease. Such systems, however, cannot deliver the level of accuracy required by an experienced physician to provide a decisive yes or no response. After conducting a comprehensive examination of the literature and collecting data from real-world case studies, a method for properly detecting cardiac anomalies is developed, which is then presented in this paper. The development of an accurate modality capable of detecting, matching patterns, forecasting, and analyzing images is required in order to provide a complimentary perspective [5, 6].

Since clinical and pathological data are the primary sources of information for diagnosing heart disease, clinicians may be unable to diagnose the condition properly in the vast majority of instances [7]. Medical practitioners may use an expert system in order to properly diagnose a patient's heart disease status and accordingly treat the patient. Data mining, combined with sophisticated hybrid algorithms, can be used to handle classification problems in medical datasets with diverse inputs. A soft computing tool for heart disease prediction and detection, in our opinion, has facilitated the development of suitably robust solutions for early treatment and prevention of heart disorders at an early stage of their development [8, 9]. Despite numerous decision support systems developed for diagnosing and predicting heart disease, researchers suggest that many remain inefficient.

This paper describes a system for heart disease prediction and classification based on machine learning and feature selection. The CFS algorithm is used to identify relevant features from the input data set, while for data categorization, the SVM RBF and random forest algorithms are used. Section 2 provides a literature review of existing techniques for preprocessing, feature selection and classification. Section 3 details the SVM RBF technique for accurate heart disease detection. Section 4 presents a result analysis of machine learning techniques used for heart disease detection. Section 5 contains the conclusion and briefly discusses future work.

2. Literature survey

Cardiovascular diseases (CVDs) remain a leading cause of mortality worldwide, necessitating the development of advanced diagnostic tools for early detection and management. Traditional methods of diagnosing heart disease, such as ECGs) and cardiac MRI, although effective, can be time-consuming and require specialized medical expertise. Recent advancements in machine learning and data mining have shown great promise in enhancing the accuracy and efficiency of heart disease prediction. By leveraging algorithms such as support vector machine (SVM) and random forest, along with feature selection techniques such as correlation-based feature selection (CFS), it is possible to develop robust models that can assist healthcare professionals in making more informed decisions. Studies have demonstrated that machine learning models can significantly improve predictive accuracy for heart disease, thereby potentially reducing the incidence of fatal outcomes [10, 11].

In more advanced applications, different types of information are bundled together into one or more categories, and then sorted according to predermined classes. When comparing datasets from the University of California, Irvine (UCI), many different approaches of characterization were used. Each individual stage of the process has been dissected in order to evaluate its precision, speed, and effectiveness. The accuracy of various classification procedures, as well as the amount of time they take to execute, have been analyzed using their respective datasets. It is clear that the performance of grouping algorithms varies depending on the dataset employed. When it comes to the effectiveness of a classifier, dataset characteristics, occurrence count, quality of features, and a number of other criteria all play a part in influencing it. J48 and Naive Bayes Updatable have both demonstrated enhanced results when used in conjunction with other informative indices [12].

Analysis of patient data in a timely manner is critical since cardiovascular coronary disease is a leading cause of mortality. Angiography, on the other hand, is a costly procedure that can have various consequences. Currently available systems gather patient data and apply a range of data mining methods in order to achieve a high level of accuracy at a lower cost and with fewer downsides. This approach has been applied to a total of 303 patient records and 54 database features [13]. The features in this knowledge base are possible indicators of coronary artery disease (CAD), according to the available treatment data [14]. The limitations are expanded, and the estimation of the degree of confidence is used as a measurement tool for predicting outcomes. This methodology offers an improved accuracy rate compared to other methods [13]. The purpose of future research is to make an educated guess about the trajectory of each individual case. It is more vital to examine how diseases impact patients than to consider the strength and characteristics of patient groups. It may be possible to leverage expanded data sets and novel architectures to produce results that are both more interesting and improved. After carrying out a number of examinations, it was found that the outcomes obtained were identical to those acquired by healthcare professionals. The calculations are put on hold while the receivers' operational attributes (ROC) curves are analyzed to determine their affectability in comparison to their explicitness. Comparisons and evaluations of the predictive capability of machine learning (ML) techniques have been used rather often and have shown that there is a need for significant improvement in this area [15].

Applying mining approaches to the information that was gathered enabled researchers to provide individuals of both sexes with answers to their inquiries about heart disease. Exploring datasets from UCI required the use of certainty as a guiding principle to analyze the data. It is more likely for men to have heart disease than it is for women. Additionally, the analysis addressed both healthy and diseased conditions with substantial quality.

Angina patients, whether male or female, who have agonizing angina or angina triggered by exertion have an increased risk of developing coronary heart disease. The slope of the resting electrocardiogram level is a crucial component in the diagnosis and assessment of various cardiac conditions. Women generally have a lower risk of coronary artery disease than men of the same age before menopause. Men, on the other hand, have a higher chance of developing CAD.

Mining tools need information scales that are standardized, clear-cut, and sometimes even multiple. It is possible to obtain various results from the same process, depending on its execution. It is essential to choose the appropriate information design for each method of characterization if one wants to get findings that can be relied upon. When making judgments and compiling clinical data, it is necessary to take into account all pertinent aspects. Absolute information serves a wide range of purposes in data mining, and it is often quite simple to use for the purpose of separating clinical information [16].

In addition to their common use in defining and illustrating foresight, decision trees may also be put to work for information mining and inductive learning.

C4.5 is a typical classification used in foresight mining; nevertheless, the accuracy of this classification decreases as problem complexity and calculations increase. Bernoulli's rule, also known as L'Hôpital's rule and first proposed by Bernoulli, aims to improve accuracy and the way the algorithm is carried out. Because of this new standard, which simplifies the process of making computational breakthroughs, futuristic judgments may now be made with a greater degree of accuracy. According to reasoning, this algorithm is effective and provides a good match for the data in applications that deal with massive volumes of data. It is essentially better in terms of its utility in the actual world. As decision tree technology develops, it is possible to obtain better decision trees, leading to more useful recommendations. When it comes to examining the business of tobacco, a number of different methods have been tried and found to be quick and effective. However, disadvantages such as high memory use rate may lead to decreased productivity as a result of larger database [17].

It is possible that dynamic calculations may be made more efficient with the use of the Bernoulli's rule, and it is also possible that the presentation of C4.5 can be improved. Utilizing the comparable rule results in a significant increase in the rate of the data collecting. The new calculation is believed to be more effective than the one that was previously used because of its capacity to process large quantities of data. It may be possible to get better recommendations by accelerating the process of developing the choice tree and acquiring an option tree with a higher level of order. Examining the transactions involving cigarettes provided support for this estimation. Without coming to a conclusion, researchers were able to obtain outcomes that were both quicker and more effective. Despite the enormous volume of data that needed to be processed, it was successful in relieving the cognitive and mental workload on individuals involved in analyzing the data [17]. However, for small datasets, the traditional C4.5 algorithm may still be more efficient than the improved version using the L'Hôpital's rule [17].

The framework primarily focuses on the evaluation and reduction of CAD. The data were examined using a methodology known as the association rule generation from feature subset procedure, which takes into account five distinct factors. The investigation required the collection of 14 different data sets from UCI before it could begin. It has been established that the method proposed by researchers is more accurate in recognizing risk and can be a decisive factor [18]. This was an important finding. In addition, mining models and rules help to lessen the negative effects of computer-aided design, which may even prevent deaths. It is important, on the other hand, to do more study in the future utilizing more significant datasets in order to determine different mining strategies and conditions [18].

Although many machine learning techniques have been used for heart disease classification and detection, there still remains a scope for research to improve accuracy, precision and recall by applying different preprocessing and feature selection methods alongside existing machine learning classifiers.

3. Methodology

In this paper, a framework for the categorization and prediction of heart disease is presented (Fig. 1). The purpose of this work is to provide a method for the prediction and classification of cardiac disease using machine learning and feature selection. The CFS method is used to extract relevant features from the input data for analysis. SVM RBF and random forest algorithms are used here for the task of data classification. Cleveland heart disease dataset is used in the experimental work. This data set has 303 instances and 14 attributes.

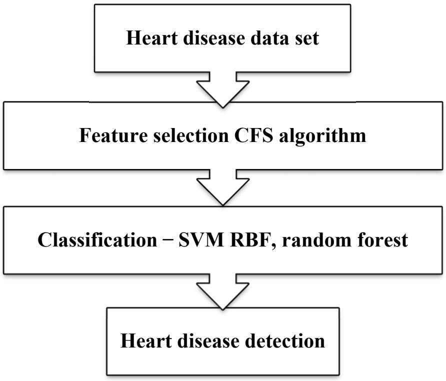
Heart disease data set
Feature selection CFS algorithm
Classification – SVM RBF, random forest
Heart disease detection

Fig. 1. Methodology for the classification and prediction of heart disease.

It is feasible to evaluate which characteristics have the most predictive potential and how much redundancy there is between them by using the CFS-based feature selection technique. In a good feature set, the features that have the greatest correlation to the class and the fewest correlations to each other are the ones that have the strongest connection to the class. The discretization of numerical data is the first step in the CFS technique since this approach only assesses the correlation between nominal features. The sole variable that may be utilized to identify features for generalized correlation-based feature selection is the correlation between any two variables. As a consequence of this, the method may be applied to a broad variety of numerical issues. CFS is an algorithm that is completely self-administered in terms of the threshold limits it imposes. Due to the fact that it makes use of the original feature space, it may be comprehended in terms of the characteristics that were already there. Because of this, the CFS filtering technique does not need a significant amount of CPU resources because the learning process is iterated over and over again [19].

The support vector machine (SVM), a non-probabilistic and binary-linear classifier, is used for the classification process. By segmenting the data into one or more training models, it is feasible to derive one or more target classes from the provided set of information. A visual representation of the data is provided in the form of the coordinates being plotted on a map. It is becoming more difficult to differentiate between various product categories as each new generation of consumer goods is released. When new instances first come into existence, such instances are categorized into certain target classes according to the side of the gap they fall on. SVM can be used to classify input datasets even if the datasets themselves are not labeled,employing the methodology of unsupervised learning in order to categorize data without predefined classes. Existing clusters, initially constructed based on functions, are expanded with the addition of new instances.

The complexity of SVM and random forest algorithms is as follows [20]:

$$\text{Training time complexity SVM} = O(n^2),$$

$$\text{Running time complexity SVM} = O(k \times d),$$

$$\text{Training time complexity random forest} = O(n \times \log(n) \times d \times k),$$

$$\text{Running time complexity random forest} = O(\text{depth of tree } \times k),$$

where n is the number of training examples, k is the number of support vectors, and d is the dimensionality of the data.

It is possible that using random forest will prove beneficial when dealing with classification and regression issues. When training decision trees, regression methods are used to anticipate the outcomes of each decision tree generated. These forecasts are used to evaluate the effectiveness of the training. When put to use in forecasting, random forest has a low standard deviation and the capacity to efficiently integrate different sets of data. Because of the unknown factors involved, the general public was first skeptical of the random forest categorization. On the other hand, it performs very well in tasks requiring prediction, as evidenced by testing [21].

4. Result analysis

This analytical work makes use of the UCI machine learning heart disease data collection [22]. This dataset has 303 instances and 14 attributes. The CFS method is applied to the input dataset in order to extract relevant features for analysis. SVM RBF and random forest algorithms are used here for data classification. Accuracy, specificity and sensitivity of RBF SVM outperform the random forest algorithm. These results are shown in Figs. 2–4.

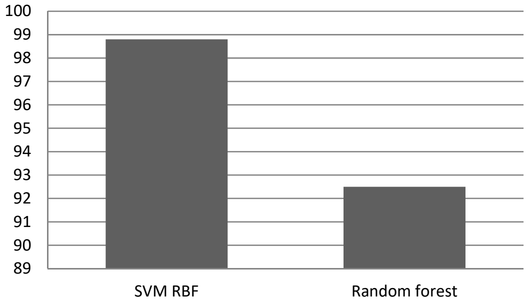

Fig. 2. Accuracy of classifiers.

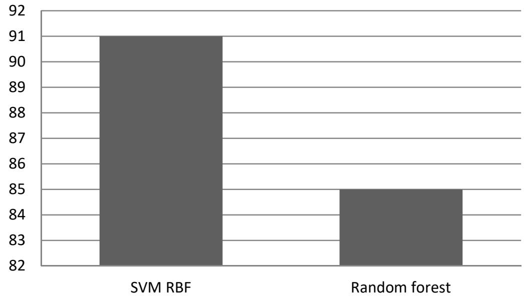

Fig. 3. Sensitivity of classifiers.

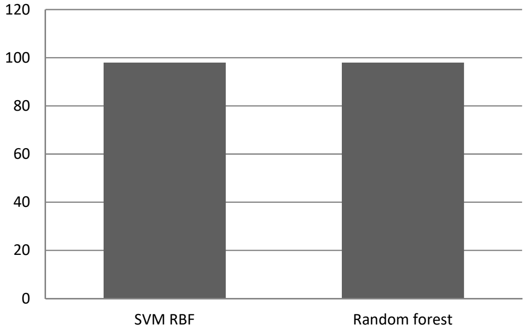

Fig. 4. Specificity of classifiers.

For result comparison, three different parameters are used:

$$\text{Accuracy} = (tp + tn)/N,$$

$$\text{Sensitivity} = tp/(tp + fn),$$

$$\text{Specificity} = tn/(tn + fp),$$

where tp, tn, fp and fn represent the sum of true positive, true negative, false positive and false negative, respectively, and N labels the total number of elements.

5. Conclusion

The heart is a very important component of the human body that should not be underappreciated or underestimated. The heart is nothing more than a pump that circulates blood throughout the body. It is possible to die within minutes after the onset of any inadequacy in body's blood circulation. The brain and other essential organs become affected, and the heart may fail. Heart-related disease specialists face an extremely tough but important task when treating patients who come for treatment of heart-related conditions. As a result, it should come as no surprise that cardiovascular disease is a leading cause of illness and mortality in modern society. An expert system with clear categorization that may assist medical professionals in identifying heart disease condition based on the clinical data of a patient is often required by physicians. This research offers a machine learning-based approach for heart disease prediction and classification, employing feature selection and classification. The CFS algorithm is used to extract relevant features from the provided dataset. SVM RBF and random forest algorithms are used for data classification. In future work, additional work can be conducted to improve the diagnostic system for identifying heart disease by incorporating optimization strategies, advanced feature selection algorithms, and classification algorithms.

References

1. A. Singh, R. Kumar, Heart disease prediction using machine learning algorithms, [in:] *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, Gorakhpur, India, pp. 452–457, 2020, doi: 10.1109/ICE348803.2020.9122958.

2. V. Sharma, S. Yadav, M. Gupta, Heart disease prediction using machine learning techniques, [in:] *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, pp. 177–181, 2020, doi: 10.1109/ICACCCN51052.2020.9362842.

3. Z.I. Attia *et al.*, An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction, *The Lancet*, **394**(10201): 861–867, 2019, doi: 10.1016/S0140-6736(19)31721-0.

4. A.E. Ulloa-Cerna *et al.*, rECHOmmend: An ECG-based machine learning approach for identifying patients at increased risk of undiagnosed structural heart disease detectable by echocardiography, *Circulation*, **146**(1): 36–47, 2022, doi: 10.1161/CIRCULATIONAHA.121.057869.

5. R.J.P. Princy, S. Parthasarathy, P.S. Hency Jose, A. Raj Lakshminarayanan, S. Jeganathan, Prediction of cardiac disease using supervised machine learning algorithms, [in:] *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 570–575, 2020, doi: 10.1109/ICICCS48265.2020.9121169.

6. M. Nasser, U.K. Yusof, Deep learning based methods for breast cancer diagnosis: A systematic review and future direction, *Diagnostics*, **13**(1): 161, 2023, doi: 10.3390/diagnostics13010161.

7. M. Sivakami, P. Prabhu, A comparative review of recent data mining techniques for prediction of cardiovascular disease from electronic health records, [in:] *Intelligent Data Communication Technologies and Internet of Things*, ICICI 2019, D. Hemanth, S. Shakya, Z. Baig [Eds.], Lecture Notes on Data Engineering and Communications Technologies, Vol. 38, pp. 477–484, Springer, Cham, 2020, doi: 10.1007/978-3-030-34080-3_54.

8. M. Sivakami, P. Prabhu, Classification of algorithms supported factual knowledge recovery from cardiac data set, *International Journal of Current Research and Review*, **13**(6): 161–166, 2021, doi: 10.31782/IJCRR.2021.13614.

9. V. Chang, V.R. Bhavani, A.Q. Xu, M.A. Hossain, An artificial intelligence model for heart disease detection using machine learning algorithms, *Healthcare Analytics*, **2**: 100016, 2022, doi: 10.1016/j.health.2022.100016.

10. A. Rajkomar *et al.*, Scalable and accurate deep learning with electronic health records, *NPJ Digital Medicine*, **1**: 18, 2018, doi: 10.1038/s41746-018-0029-1.

11. J.M. Kwon *et al.*, Deep learning-based algorithm for detecting aortic stenosis using electrocardiography, *Journal of the American Heart Association*, **9**(7): e14717, 2020, doi: 10.1161/jaha.119.014717.

12. A. Frank, A. Asuncion, UCI Machine Learning Repository, University of California, School of Information and Computer Sciences, Irvine, 2010, http://archive.ics.uci.edu/ml.

13. A. Ghorbani *et al.*, Deep learning interpretation of echocardiograms, *NPJ Digital Medicine*, **3**: 10, 2020, doi: 10.1038/s41746-019-0216-8.

14. V. Parthasarthy, L. Pallavi, H. Anandaram, M. Praveen, S. Arun, R. Krishnamoorthy, Prediction of coronary artery disease by adapting hybrid approach of machine learning methods, [in:] *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, pp. 1233–1237, 2022, doi: 10.1109/ICOSEC54921.2022.9952111.

15. J. N., D. P., M. E., R. Santhosh, R. Reshma, D. Selvapandian, Heart attack prediction using machine learning, [in:] *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, pp. 854–860, 2022, doi: 10.1109/ICIRCA54612.2022.9985736.

16. J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, 2011.

17. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

18. M.A. Jabbar, P. Chandra, B.L. Deekshatulu, Prediction of risk score for heart disease using associative classification and hybrid feature subset selection, [in:] *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, Kochi, India, pp. 628–634, 2012, doi: 10.1109/ISDA.2012.6416610.

19. A. Meena Kowshalya, R. Madhumathi, N. Gopika, Correlation based feature selection algorithms for varying datasets of different dimensionality, *Wireless Personal Communication*, **108**: 1977–1993, 2019, doi: 10.1007/s11277-019-06504-w.

20. O. Chapelle, Training a support vector machine in the primal, *Neural Computation*, **19**(5): 1155–1178, 2007, doi: 10.1162/neco.2007.19.5.1155.

21. C.-H. Weng, T.C.-K. Huang, R.-P. Han, Disease prediction with different types of neural network classifiers, *Telematics and Informatics*, **33**(2): 277–292, 2016, doi: 10.1016/j.tele.2015.08.006.

22. A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, Heart Disease. UCI Machine Learning Repository, 1988, doi: 10.24432/C52P4X.

Received April 30, 2022; revised version October 7, 2022;
accepted October 12, 2022; published online August 13, 2024.