A Comprehensive Review of Various Machine Learning Techniques for Heart Disease Prediction

Guna Sekhar Sajja
Information Technology
University of the Cumberlands
Kentucky, USA

ABSTRACT

Data mining techniques have been used by several researchers to detect illnesses. Some methods are intended to predict a single sickness, while others are intended to predict a wide variety of diseases. It is also possible to improve the accuracy of sickness prediction. In this post, we provided an overview of data classification approaches that are available. These algorithms are mostly represented by themselves. The classification of data is a common and computationally difficult procedure. We've also established the foundation for data categorization. We would compare the best algorithms from a huge set of existing algorithms. This article presents a summary of the research on machine learning and soft computing-based methods for categorizing and predicting cardiac disease.

Keywords

Data Mining, Machine Learning, Classification, Prediction, Heart Disease

1. INTRODUCTION

Approximately 5% of people under the age of 35 in developed countries and over 20% of those over the age of 75 suffer from heart disease, which is a dangerous condition. Hospital admissions are affected by heart collapse at a rate of 3–5%. Heart failure is the most common reason for hospitalization, according to clinicians in clinical practice. Up to 20% of total health expenditure in wealthy countries is devoted to this issue.

The heart has several internal organs that might be affected by different types of cardiac illness. Heart problems are therefore classified as cardiovascular illnesses [3], and some of these conditions are discussed in greater detail below. In the world, the most common type of cardiac illness is coronary artery disease (CAD). Coronary artery disease is another name for it (CAD). It's a condition marked by fat buildup in the blood vessels and capillaries. A blockage in the heart's veins and capillaries means that the organs inside the heart don't get as much oxygen or blood as they need. It is also shown in figure 1.

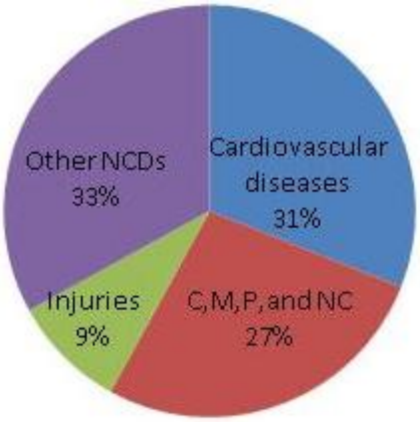

Figure 1: Major Causes of Death

Classifier model of progressing feature inclusion with back-elimination discovered by the authors of [1] on multiple datasets, including Arrhythmia, heart disease and ECG datasets. Experimental findings show that the feature selections enhanced categorization operations while reducing inputs. According to this research, using the supplied arrhythmia dataset improves results by 78% while lowering features by 19%. There was an 85% increase in performance, and the number of features was reduced from eight to four. Reductt characteristics have been shown in earlier studies to improve classifier performance.

An ANN-based fuzzy inference system described in [2] employs ANFIS and linear LDA approaches to provide a predictive solution for damaging information and forecast. Because of this hybrid classification strategy, the inference system outperforms conventional methods in terms of efficiency and performance. Early detection and prevention of coronary heart disease are made possible thanks to the use of this blood test.

Sickness management requires extensive data analysis before a plan can be developed. Early illness detection with artificial intelligence, assessing severity, and making early predictions are the most typical occurrences. This will help slow the disease's course, enhance the lives of sufferers, and save money on medical expenses. Approaches based on machine learning have been employed in this area.

2. LITERATURE REVIEW

Individual neural networks in an ensemble model may be trained by using a cooperative neural network ensemble, as demonstrated by a researcher in [3]. (CNNEs). The accuracy metrics used to discover hidden nodes in a single neural network are improved using this model's constructive technique. Using negative correlation theories, neural networks in a multi-layered model are gradually trained. It is possible to maintain variety over several layers of neural networks by using negative correlation learning methods. Datasets for cancer diagnosis, diabetes prediction, heart disease prediction and letter recognition have all been used extensively to test this method. The results show that it creates a neural network model with a greater generalization ability.

A swarm optimization methodology-based fuzzy expert system is described [4]. UCI machinery cardiac data sets were largely the focus of this system's creation. A decision tree-based algorithm is used in this approach to discover key characteristics that help with accurate diagnosis and treatment. The output data is generated using a fuzzy rule foundation. The outcome is derived via fuzzy approximation. The final fuzzy expert system has a precision of 93.27 percent thanks to the particle swarm optimization approach. Unlike other categorization approaches, fuzzy expert systems provide an output model that is simple to understand when compared to others.

The authors [5] suggested a firefly-based algorithm using rough sets for making reliable predictions. Using fuzziest and harshest theoretical concepts together lowers the problems in heart disease datasets produced by uncertainties and high-dimensional variables. It is possible to obtain the best answers with the least amount of computing operations by using the roughest-based fuzzy learning method When it comes to heart disease prognosis and treatment, the results of this study outperform those of support vector machines and artificial neural networks (ANNs).

Researchers have devised a method for forecasting ventricular arrhythmias [6]. This paper describes a pain prediction system that utilizes an ECG signal processor incorporated into the system. Ventricular arrhythmia can be predicted in people using a specific set of ECG features. There is an ECG wave detection and marking (PQRST) function that tracks the fiducial location. There are real-time and adaptive approaches used throughout this operation. ECG signal fluctuations may be successfully controlled using these technologies, resulting in readings that are highly sensitive and accurate. Using cardiovascular signals from the American Heart Association database, the system's performance is assessed. As a result of the simulations, we now have accuracy measures that are on par with those of previous approaches. The simulation is carried out using an IC tailored to the individual application (ASIC). For the record, this is the first time ESP has been used in an ASIC for the prediction of ventricular arrhythmia.

Researchers [7] presented a method for forecasting adult cardiovascular disease risk based on a naive Bayes classifier. To begin, it examines medical records in order to establish whether or not the patient has a cardiac condition. Our goal is to increase the sensitivity, accuracy, and specificity of heart disease categorization and prediction by implementing this new technology. Primary risk factors for heart disease such as diabetes, blood cholesterol levels, and renal and vascular function are targeted to catch risk factors early. It divides danger into three levels: level 1, level 2, and level 3. Research shows that this strategy is superior to others for predicting heart disease (more than 80 percent). This study accurately and systematically predicts heart illness as proven by patients, cardiologists, and other medical experts, too.

We present an effective model-based recommendation system that employs machine learning techniques in this study [8]. Using machine learning algorithms and quick Fourier transform technology, it anticipates heart issues and provides medical suggestions to patients. During time series decomposition, the frequency of the incoming data is captured using quick Fourier transformation methods. Using simple learning methods, the patient's health data are anticipated sooner, and suggestions are provided ahead of time. These strategies rely on real-time heart disease datasets collected from cardiac patients. According to the data, this strategy improves prediction accuracy while demanding less computation work from patients. It is also a useful tool for providing speedy analysis and therapy recommendations to cardiac patients.

Researchers [9] revealed a naive Bayes classification method for diagnosing cardiac problems. This book focuses on the repercussions of heart disease in today's society. It combines statistical methodologies with the Nave Bayes classifier to predict and diagnose heart diseases effectively. It employs data pre-processing techniques to successfully deal with the massive and difficult collection of medical data. To classify various forms of heart illness, a discretization method is applied. The discretization approach used in this case is supervised discretization with equal frequencies. The project makes use of heart disease datasets from the stat log heart database. According to the results, it outperforms earlier techniques in terms of accuracy.

Authors [10] explains the linear SVM classification model. After isolating a hyperplane from each provided dataset using labeled training samples, this discriminative classifier provides an ideal hyperplane. This further categorizes the new instances of the input data model. A hyperplane in two dimensions is a line that divides the specified hyperplane into two half. Each class is on either side of the dividers. SVM delivers class separation in a nutshell.

The linear combinational model for identifying the optimal hyperplane was described by the authors [11]. It determines the optimum hyperplane for classifying the target classes. The hyperplane should be chosen in such a way that it classifies the target class efficiently. Depending on the situation, the hyperplane is chosen differently. If the target classes are accurately separated by three hyperplanes. In this case, the concept of distance margin is used. It makes it simple to locate the appropriate hyperplane. Similarly, the margin distance measurements are adjusted based on the circumstances to provide better results. The maximum margin classifier and the soft margin classifier are two SVM classification techniques that use margin distance measurements. This is used to generate an efficient categorizing technique.

The authors [12] published a new paper that included a deep learning based linear SVM classification. It replaces the soft-max layer in deep convolutional networks with a linear collaborative machine. This strategy reduces margin-based loss as an alternative to cross-entropy loss. There is research that shows how the SVM can be used to replace different layers of a deep convolutional network. In this work, the SVM is replaced by the second layer of the deep convolution network. The principal application of this research is human face recognition systems.

An SVM-based technique for diagnosing coronary artery disease was described by the authors [13]. To predict and diagnose coronary artery disease, this study combines a support vector machine with principal component analysis

(PCA). The experimental procedure was applied to 480 patients, each of whom had 23 features. To provide higher performance metrics, principal component analysis minimizes the depth of the characteristics in the experimental dataset. An optimal SVM model for each reduced dimension is determined through attribute reduction. According to the results of the experiments, this strategy minimizes training error while needing less training and testing time.

A strategy for acute diagnosis and prognosis of coronary diseases is described in this work [14]. The primary purpose of the research is to identify heart abnormalities and help medical experts discover and treat them early. An SVM classifier is used in this work to evaluate patient data such as age, gender, and risk characteristics in order to provide better medical suggestions. The system's performance is evaluated using laboratory and clinical data from around 228 people. Four different approaches were tested on the same dataset. The SVM classifier is more accurate in predicting the diagnosis of heart disease. Among the 228 patients whose samples were evaluated, 99 predictions were revealed. This study clearly illustrates that implementing machine learning techniques throughout healthcare systems greatly assists medical practitioners in making clear and concrete decisions and diagnoses.

For medical anesthetic, the authors [15] presented neural network ensemble models. A number of estimation criteria are used to test the presentation of these classifiers using real-world datasets. A comparison is made between the basic classifier and the targeted classifier in order to increase performance.

The authors [16] presented an Atrial arrhythmias modification technique for distinguishing between critical locations inside the major atrial bundles. Extremely rapid critical movement in the atria can imitate distorted P-wave morphology (PWM) on the ECG. The method proved realistic, predicting the atrial position with an accuracy of 85%.

Researchers [17] investigate the challenges and potential solutions for establishing a clinical decision support system for the prediction and diagnosis of cardiac disease. This project is mostly focused with cardiovascular issues in Iranian society.

A combination of support vector machine and binary particle swarm optimization is utilized for prediction and classification. The binary particle swarm optimization method aids in feature selection. SVM is the major player here, and it is inextricably linked to the data classification process.

he Isfahan Healthy Heart Program (IHHP) dataset is used to test system performance measurements, and it gives enhanced accuracy, sensitivity, and specificity values when compared to commonly used classifiers for heart disease prediction. Support vector machines are the most successful classification tool, and they are widely employed in a variety of disciplines. It also has various uses in the prediction and management of cardiac disease. This approach works well with the precise margin of separation and is much more successful in high-dimensional areas. It also operates effectively with a greater number of dimensions than samples. Furthermore, this approach is both computationally and memory efficient. When dealing with noisy and ambiguous data, performance measurements can suffer.

Authors [18] examine the statistical approach in depth, proving its applicability to the logistic regression data analysis model. Logistic regression is a statistical data analysis method that works with binary dependent variables. The logistic regression methodology is a powerful tool for performing regression analysis. The logistic regression techniques are used to estimate the logistic model's parameters. The probability of an occurrence in a logistic model is defined as a linear combination of independent elements.

A comparison of many categorization algorithms for the prediction of coronary heart disease is offered in this work [19]. The performance measures of numerous algorithms used to forecast coronary heart disease are examined in this proposal. The analysis was performed on 1245 user items, 865 of which had heart problems and 380 of which did not. Techniques such as logistic regression, CART, multilayer perceptron, radial basis function, and self-organizing feature maps are used in the analysis process. Predictive factors include age, gender, and diabetes status. Performance measurements such as hierarchical cluster analysis and multidimensional scaling are used to determine the classification algorithms' performance. The logistic regression model outperforms the other alternatives, as evidenced by its superior performance. The multilayer perceptron, on the other hand, provides the initial performance measurements.

Authors in [20] describe an approach for estimating the risk of cardiovascular disease. It overcomes the fundamental drawback of existing categorization methods. It also investigates the feasibility of forecasting heart illness using existing algorithms. The primary focus of this research is on reclassification tables, as well as classification process sensitivity and specificity measures. The application of these measurements throughout this work improves the accuracy of heart disease classification. It leverages logistic regression models in the categorizing process. According to the results of the experiments, it enhances accuracy and performance measures. It investigates the progression of atherosclerosis and identifies the primary risk factors for rheumatoid arthritis. Carotid ultra sonography is used to identify risk factors for cardiovascular disease. Logistic regression models are used to separate the ultrasonic readings and develop the classification procedure. Furthermore, the regression models correctly identify the baseline variables and produce a predictive regression. In terms of performance measures, this strategy outperforms the other traditional approaches, according to the experimental data.

Logistic regression models are the most reliable classification method and are widely used in a wide range of applications. This method is typically used to evaluate the performance of cardiac disease prediction systems. This method avoids overfitting and yields more accurate results. However, dealing with nonlinear relationships is more challenging and time intensive. Furthermore, rather of a classification model, this technique is effective as a performance evaluation tool in healthcare organizations.

The authors [21] presented a hybrid A-priori technique that combines k-means and the Apriori algorithm. The dataset is first clustered using the k-means clustering algorithm. In the following phase, the apriori approach is utilized to discover the frequent itemset. The Boolean association rule uses a "bottom-up" method to achieve a better result. The patterns oversee responding to the complex questions posed by the real-world scenario of the heart disease prediction system. Classification is the method in predictive analysis that correctly classifies the input data and maps it to the appropriate classes. Data is classified into two types: labeled data and unlabeled data. The labeled data contains several predictor traits as well as a single target property. Each value of the target characteristics represents the class label. In the unlabeled traits, the predictor features are the only ones. The classification process's primary purpose is to correctly predict the class of unlabeled data using classification models created from labeled samples (historical data). To begin, a training model with known related class (or target values) is built.

3. CONCLUSION

The most prevalent reason for admission seen by clinicians in their clinical practice is heart failure. The costs are significant, accounting for up to 20% of total health expenditure in affluent countries. This article gives an in-depth examination of several machine learning and soft computing-based strategies for heart disease prediction. A variety of disorders can be forecasted using data mining algorithms based on decision trees. This study's findings will have a big impact in the medical field. It has the potential to be tremendously beneficial to both patients and clinicians. Further work should be done to increase the classifier's accuracy using other data mining classification methods such as rule-based inference and expectation maximization.

4. REFERENCES

[1] Shilaskar, S. and Ghatol, A. (2013), ‘Expert systems with applications feature selection for medical diagnosis : evaluation for cardiovascular Diseases’, Journal of expert sy stem with application 4(10), 4146–4153.

[2] Yang, X., Li, M., Zhang, Y. and Ning, J. (2014), ‘Cost-sensitive naive bayes classification of uncertain data’, Journal of Scientific World 9(8), 1897–1904.

[3] Islam, Monirul, Y. X. and Murase (2003), ‘A constructive algorithm for training cooperative neural network ensembles’, Journal of IEEE transactions on Neural Networks 14(4), 820–34.

[4] Muthukaruppan, J. (2012), ‘A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease’, Journal of Expert Systems With Applications 9(4), 11657–11665.

[5] Long, Nguyen Cong, M. H. (2015), ‘A highly accurate firefly based algorithm for heart disease prediction’, Journal of Expert Systems with Applications 42(21), 8221–8231.

[6] Bayasi, N. and Tekeste (2016), ‘Low-power ECG-based processor for predicting ventricular arrhythmia’, Journal of IEEE transactions on very large scale integration systems 24(5), 1962–1974.

[7] Liu, Wang, M., Moran, A. E., Liu, J. and Coxson (2016), ‘Projected impact of salt restriction on prevention of cardiovascular disease in china: a modeling study’, Journal of plos one 11(2), 1–16.

[8] Zhang, Shuai, Y.-L. S. A. (2017), ‘Deep learning based recommender system: a survey and new perspectives’, Journal of ACM Computing Surveys 1(1), 1–35.

[9] Aydin, S. (2016), ‘Comparison and evaluation data mining techniques in the diagnosis of heart disease’, Indian journal of science and technology 6(1), 420–423.

[10] Berikol, B. and Yildiz (2016), ‘Diagnosis of acute coronary syndrome with a support vector machine’, Journal of Medical System 40(4), 11–18.

[11] AminKhatami, AbbasKhosravi, C. L. (2017), ‘Medical image analysis using wavelet transform and deep belief networks’, Journal of Expert Systems With Applications 3(4), 190–198.

[12] Chang and Lin (2013), ‘A library for support vector machines’, Journal of ACM transactions on intelligent systems and technology 5(39), 724–749.

[13] Hannan, S. A. and Bhagile (2010), ‘Diagnosis and medical prescription of heart disease using support vector machine and feedforward backpropagation Technique’, International Journal on Computer Science and Engineering 02(06), 2150–2159.

[14] Berikol, B. and Yildiz (2016), ‘Diagnosis of acute coronary syndrome with a support vector machine’, Journal of Medical System 40(4), 11–18.

[15] Cheng-Hsiung Wenga, Tony Cheng-Kui Huang, R.-P. H. (2016), ‘Disease prediction with different types of neural network classifiers’, Journal of Telematics and Informatics (4), 277–292.

[16] Vafaie, Ataei, M. (2014), ‘Heart diseases prediction based on ECG signals classification using a genetic-fuzzy system’, Journal of biomedical signal processing and control 14(5), 291–296.

[17] Sali, R. and Shavandi, M. (2016), ‘A clinical decision support system based on support vector machine and binary particle swarm optimisation for Cardiovascular disease diagnosis’, International Journal of Data mining and Bio-informatics 15(1), 312–327.

[18] Hssina, Merbouha, A. and Ezzikouri (2014), ‘A comparative study of decision tree ID3 and C4.5’, International Journal of Advanced Computer Science and Applications 4(2), 13–19.

[19] Kurt, Imran, U. M. (2008), ‘Comparing performances of logistic regression , classification and regression trees and artificial neural networks for predicting albuminuria in type-2 diabetes mellitus’, Journal of Expert System with Applications 31(10), 173–187.

[20] Pencina, M. J. and Derpan (2010), ‘Heart disease diagnosis using machine learning approach’, Journal of NIH public access 119(24), 3078–3084.

[21] Sairabi, Mujawar, D. (2015), ‘Prediction of Hear t Disease using Modified K-means and by using Naive Bayes’, International Journal of Innovative Research in Computer and Communication Engineering 3.