International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 6 (2016) pp 4377-4380

Study on Cardiovascular Disease Classification Using Machine Learning Approaches

R. Subha
Assistant Professor, PSG College of Arts and Science, Coimbatore, Tamil Nadu, India.

K. Anandakumar
Assistant Professor, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India.

A. Bharathi
Assistant Professor, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India.

Abstract
The diagnosis of heart disease which depends in most cases on complex grouping of clinical and pathological data. Due to this complexity, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart diseaseprediction. In case of heart disease, the correct diagnosis in early stage is important as time is very crucial. Numeral number of tests must be requisite from the patient for detecting a disease. Machine learning based method is used to classify between healthy people and people with disease. Cardiovascular disease is the principal source of deaths widespread and the prediction of Heart Disease is significant at an untimely phase. In order to reduce number of deaths from heart diseases there has to be a quick and efficient detection technique. This work presents a comprehensive review for the prediction of cardiovascular disease by using machine learning based approaches.

Keywords*:* Cardiovascular disease, classification, machine learning

Introduction

A broad spectrum of disorders is Cardio Vascular Disease (CVD) which is a label for affecting both the heart muscle itself (i. e. myocardial infarction) and the vasculature (i. e. hypertension). In the western world, cardiovascular disease will remain the morality cause and it is cause for deaths more than 16 million annually. By changing the style of life reducing the cholesterol intake and regularly exercising will decrease the fatal event chances linked with CVD. But, the critical step is that early detection of CVD for preventing the death associated with CVD. So, an adequate visit to doctor will results in large volumes of data's of patient which includes ECG is an important step towards early detection which must be examined carefully.

In order to assist the medical professionals, medical diagnostic based on computer have been developed for analyzing the large volumes of the patient data. These system efficacy mainly depends on the features which are used must be correlated with some disease state. Based on ECG signals, several signal processing techniques have been implemented successfully which will extract a set of features which is used subsequently by many machine learning classification tools. The purpose of the review is to assess the evidence of healthcare benefits involving the application of machine learning to the clinical functions of diagnosis and analysis.

Many researchers are interested in using classification method for clinical research nowadays. Accurate classification of disease states (disease present/absent) or of disease etiology or subtype allows subsequent investigations, treatments, and interventions to be delivered in an efficient and targeted manner. Similarly, accurate classification of disease states permits more accurate assessment of patient prognosis. This works provides the classification method for classifying the CVD patients accurately.

Literature Survey

A survey on present techniques using data mining technique is provided by Soni *et al* (2011) which are used in the current medical research mainly in prediction of heart disease. A method is developed by Beigi *et al* (2011) for predicting the proteins based on the derived features from the Chou's pseudo amino acid composition server and a powerful machine learning approach Support Vector Machine (SVM) is used. For the evaluation and construction of probability and classification estimation rules are described by Kruppa *et al* (2012). They reviewed and explained the machine learning approach in details. Austin *et al* (2013) developed the alternate classification schemes based on the machine learning literature and data-mining which includes bootstrap aggregation (bagging), random forests, boosting and support vector machines. Rotation Forest (RF) is constructed by Ozcift and Arif (2011) for evaluating their classification performances using heart diseases, Parkinson's and diabetes from the literature.

The classification technology is applied by Yeh *et al* (2011) for constructing an optimum prediction model of cerebrovascular disease. This model will extract and improve the prediction and diagnosis of cerebrovascular disease. The artificial neural network is evaluated by Al-shayea (2011) for

diagnosing the disease. A GUI based interface is designed by Soni*et al* (2011) for entering the record of the patient and predicting whether the patient is having the heart disease or not using the classifier based on weighted association rule. Anooj (2012) developed machine learning techniques for gaining the knowledge automatically from raw data or examples. Parthiban *et al* (2011) aims to predict the diabetes patient chances of getting heart disease using Naïve Bayes data mining classifier which will produce an prediction model using minimum training set.

A method is presented by Khalilia*et al* (2011) by utilizing the utilization projectCost and Utilization Project (HCUP) dataset to predict the individual risk of disease based on their medical history. A computational intelligence technique is investigated by Nahar *et al* (2013) for detecting the heart disease. They highlighted the expert judgment based feature selection process and compared against the generally computational intelligence which based on feature selection mechanism. Using ensemble based methods, Austin *et al* (2012) evaluated the improvements achieved which include random forests, bootstrap aggregation of regression trees and boosted regression trees. Kumari and Sunila (2011) analyzed the data mining classification techniques such as decision trees, artificial neural network, SVM and RIPPER classifier for diagnosing the cardiovascular disease dataset. CVD is associated with cardio respiratory fitness morality. Cox proportional hazards models were used to estimated by Gupta *et al* (2011) the risk of CVD mortality with a traditional risk factor model (age, sex, systolic blood pressure, diabetes mellitus, total cholesterol, and smoking) with and without the addition of fitness.

Comparative Study

AUTHOR	DESCRIPTION	PROS AND CONS
Soni *et al* (2011)	A survey on present techniques using data mining technique is provided which are used in the current medical research mainly in prediction of heart disease.	The main advantage is that the accuracy of the decision tree and bayesian classification will be improved and reduced the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. But this work can be further enhanced and expanded for the automation of heart disease prediction.
Beigi *et al* (2011)	A method is developed for predicting the proteins based on the derived features from the Chou's pseudo amino acid composition server and a powerful machine learning approach Support Vector Machine (SVM) is used.	The advantage is that the method is able to predict two major subclasses of MMP family; Furin-activated secreted MMPs and Type II trans-membrane.
Kruppa *et al* (2012)	For the evaluation and construction of probability and classification estimation rules are described. They reviewed and explained the machine learning approach in details	This study provides the construction and evaluation of classification and probability estimation rules which is the main benefit.
Austin *et al* (2013)	They developed the alternate classification schemes based on the machine learning literature and data-mining which includes bootstrap aggregation (bagging), random forests, boosting and support vector machines.	Modern data mining and machine learning methods offer advantages to predict the heart failure patient based on subtype
Ozcift and Arif (2011)	Rotation Forest (RF) is constructed for evaluating their classification performances using heart diseases, Parkinson's and diabetes from the literature	Machine learning applications, particularly CADx systems, needs classifiers with enhanced accuracies. But this study did not evaluate the effect of feature selection algorithm on classifier performances
Yeh *et al* (2011)	The classification technology is applied for constructing an optimum prediction model of cerebrovascular disease. This model will extract and improve the prediction and diagnosis of cerebrovascular disease	Eight important influence factors of predicting cerebrovascular disease and 16 diagnosis classification rules were extracted which is the main advantage.
Al-Shayea(2011)	The artificial neural network is evaluated for diagnosing the disease.	The main benefit is the use of a hybrid model for adapting the content according to the style of the learner, and also in the maintenance of style via the trace routes and the degree of adaptation.
Soni*et al* (2011)	A GUI based interface is designed for entering the record of the patient and predicting whether the patient is having the	A GUI has been designed to enter the patient's records and the presence of heart disease for the patient

	heart disease or not using the classifier based on weighted association rule	is predicted using the rules stored in the rule base.
Anooj (2012)	They developed machine learning techniques for gaining the knowledge automatically from raw data or examples.	The automatic procedure to generate the fuzzy rules is an advantage of the proposed system and the weighted procedure introduced in the proposed work is additional advantage for effective learning of the fuzzy system.
Parthiban*et al* (2011)	They aims to predict the diabetes patient chances of getting heart disease using Naïve Bayes data mining classifier which will produce an prediction model using minimum training set	The advantage is that predicting the chances of getting a heart disease using attributes from diabetic's diagnosis.
Khaliliа*et al* (2011)	A method is presented by utilizing the utilization projectCost and Utilization Project (HCUP) dataset to predict the individual risk of disease based on their medical history.	Extensive proof that RF can be successfully used for disease prediction in conjunction with the HCUP dataset which is the main advantage.
Nahar *et al* (2013)	A computational intelligence technique is investigated for detecting the heart disease. They highlighted the expert judgment based feature selection process and compared against the generally computational intelligence which based on feature selection mechanism.	The advantages are the best suited algorithms for heart disease diagnosis is identified; and Medical feature selections combined with the computerized feature selection (CFS) are considered.
Austin *et al* (2012)	Using ensemble based methods, they evaluated the improvements achieved which include random forests, bootstrap aggregation of regression trees and boosted regression trees.	Ensemble methods from the data mining and machine learning literature increase the predictive performance of regression trees.
Kumari and Sunila (2011)	They analyzed the data mining classification techniques such as decision trees, artificial neural network, SVM and RIPPER classifier for diagnosing the cardiovascular disease dataset	The main advantage is the RIPPER classifier is used in this research.
Gupta *et al* (2011)	CVD is associated with cardiorespiratory fitness morality. Cox proportional hazards models were used to estimated the risk of CVD mortality with a traditional risk factor model (age, sex, systolic blood pressure, diabetes mellitus, total cholesterol, and smoking) with and without the addition of fitness	The addition of fitness to traditional risk factors significantly improves reclassification of the risk of CVD mortality across short-term and long-term follow-up.

The above table shows the pros and cons of several literatures, different authors give various ideas to identify the cardiovascular disease, but there is no effective method. This motivates the new approach for identifying the cardiovascular disease.

Problems and Directions

Diagnosis is clearly a difficulty in cardiovascular disease, and an effective screening process, particularly one that doesn't require a clinic visit, would be beneficial. Mainly the heart attacks will be occurred due to the plaque in the clot and artery ruptures and then forms which stops the flow of blood. And the heart disease diagnosis is based on the knowledge given only by the patients. So, several problems will occur for proper diagnosis of the heart patient. One complication that may exist in available medical data is an inconsistency across datasets. For example, multiple datasets for a given disorder often exist, collected from different sources and using slightly different features. Combining them in some effective way into a large, cohesive dataset would result in a more robust and well-trained learner. Another complication is an occasional lack of labeled examples. Austin et al (2013) study had a very limited focus: comparing the ability of different methods to predict or classify disease subtype in patients hospitalized with HF in Ontario. However, conventional logistic regression was able to more accurately predict the probability of the presence of HFPEF amongst patients with HF compared to the methods proposed in the data mining and machine learning literature. There is limited research comparing the performance of different classification/prediction methods for predicting the presence of disease, disease etiology, or disease subtype. A successful machine learning approach to classification would be applicable to many types of medical diagnosis.

Conclusion

This survey provides the brief description of machine learning techniques for classification of cardiovascular disease. The classification accuracy depends on the exact metrics which are used which also indicates the variety of features has been utilized. But, there is no systematic study performed for the accurate prediction. The role of classifier is important in the clinical setting so that the results can be used for determining the treatment. The researchers will generate the results required which are considered important if these classifiers are vital part of medical arsenal. Only then can these exciting results engender a multi-disciplined approach to medical research.

References

[1] Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive data mining for medical diagnosis: An overview of heart disease prediction. "*International Journal of Computer Applications* 17, no. 8 (2011): 43-48.

[2] Beigi, Majid Mohammad, Mohaddeseh Behjati, and Hassan Mohabatkar. "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. " *Journal of Structural and Functional Genomics* 12, no. 4 (2011): 191-197.

[3] Kruppa, Jochen, Andreas Ziegler, and Inke R. König. "Risk estimation and risk prediction using machine-learning methods. " *Human genetics* 131, no. 10 (2012): 1639-1654.

[4] Austin, Peter C., Jack V. Tu, Jennifer E. Ho, Daniel Levy, and Douglas S. Lee. "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. " *Journal of clinical epidemiology* 66, no. 4 (2013): 398-407.

[5] Ozcift, Akin, and Arif Gulten. "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. " *Computer methods and programs in biomedicine* 104, no. 3 (2011): 443-451.

[6] Yeh, Duen-Yian, Ching-Hsue Cheng, and Yen-Wen Chen. "A predictive model for cerebrovascular disease using data mining. " *Expert Systems with Applications* 38, no. 7 (2011): 8970-8977.

[7] Al-Shayea, Qeethara Kadhim. "Artificial neural networks in medical diagnosis. "*International Journal of Computer Science Issues* 8, no. 2 (2011): 150-154.

[8] Soni, Jyoti, Uzma Ansari, Dipesh Sharma, and Sunita Soni. "Intelligent and effective heart disease prediction system using weighted associative classifiers. " *International Journal on Computer Science and Engineering* 3, no. 6 (2011): 2385-2392.

[9] Anooj, P. K. "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. " *Journal of King Saud University-Computer and Information Sciences* 24, no. 1 (2012): 27-40.

[10] Parthiban, G., A. Rajesh, and S. K. Srivatsa. "Diagnosis of heart disease for diabetic patients using naive bayes method. " *International Journal of Computer Applications* 24, no. 3 (2011): 7-11.

[11] Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. "Predicting disease risks from highly imbalanced data using random forest. " *BMC medical informatics and decision making* 11, no. 1 (2011): 51.

[12] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. " *Expert Systems with Applications* 40, no. 1 (2013): 96-104.

[13] Austin, Peter C., Douglas S. Lee, Ewout W. Steyerberg, and Jack V. Tu. "Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods?. "*Biometrical journal* 54, no. 5 (2012): 657-673.

[14] Kumari, Milan, and Sunila Godara. "Comparative study of data mining classification methods in cardiovascular disease prediction 1. " (2011).

[15] Gupta, Sachin, Anand Rohatgi, Colby R. Ayers, Benjamin L. Willis, William L. Haskell, Amit Khera, Mark H. Drazner, James A. de Lemos, and Jarett D. Berry. "Cardiorespiratory fitness and classification of risk of cardiovascular disease mortality. " *Circulation* 123, no. 13 (2011): 1377-1383.