Volume 7, Issue 4, April 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining, Soft Computing, Machine Learning and Bio-Inspired Computing for Heart Disease Classification/ Prediction – A Review

M. Rathi*, B. Narasimhan
Assistant Professor, Department of Computer Technology, Dr. N.G.P. Arts and Science College, Coimbatore, Tamilnadu, India

Abstract: ***Data mining is the most common research area in the field of computer science and allied areas. Decision making in clinical data mining plays a significant role in patient's life. In this survey research article we aim to portray various data mining algorithms, soft computing techniques, machine learning algorithms and bio-inspired algorithms for predicting / classifying heart disease. Several mechanisms namely apriori algorithm, frequent itemset mining, support vector machine, neural network, classification and regression trees, fuzzy rule-based clinical decision support system, k-nearest neighbor, genetic algorithm, scoring system, nature language processing (NLP) techniques, type-2 fuzzy logic system, decision tree and statistical methods are used to classify heart disease prediction. From this survey research, it is identified that support vector machine algorithm outperforms all the other methods in terms of classification accuracy.***

Keywords: ***Data mining, clinical decision support system, predictive data mining, classification, heart disease, soft computing, machine learning, optimization.***

I. INTRODUCTION

Data mining is a presentation computational method in order to obtain hidden knowledge from data warehouses and also from large databases. Several data mining mechanisms / algorithms are used in recent years for defining novel models for heart disease prediction / classification. In 2020, as per the World Health Organizations' (WHO) report more than 70% of deaths among world population will be due to heart diseases. There are various types of heart diseases namely Coronary Artery Heart Disease (CAHD), Ischemic Heart Disease (IHD), Cardiac Arrest (CA) and even more. The conventional hard computing techniques are not far enough to classify or predict such heart diseases. Hence several research works are carried out using conventional data mining algorithms, soft computing techniques, machine learning algorithms and bio-inspired mechanisms. Soft computing is a fusion of research in evolutionary algorithms and genetic programming, neural science and neural net systems, fuzzy set theory and fuzzy systems, and chaos theory and chaotic systems. Another crucial research issue falls in finding the best solution for identifying the complexity of computations. These research trends tend to offer a wide scope of research in predicting / classifying heart diseases. This paper is organized as follows. This section provides a descriptive report of heart disease prediction using several mechanisms. The next section discusses on the related works carried out. Section 3 describes about performance metrics used. Section 4 projects findings and conclusions.

II. RELATED WORKS

Nahar et al. [10] investigated the sick and healthy factors which contribute to heart disease for males and females. To identify these factors, the author had used an approach called Association rule mining with UCI Cleveland dataset with the three rule generation algorithms like Apriori, Predictive Apriori and Tertius. Female were seen to have less chance of coronary heart disease than male after the analysis of factors on sick and healthy individuals and taking confidence as an indicator. In addition, the attributes indicating healthy and sick conditions were also identified. The author had also stated that factors such as chest pain being asymptomatic and the presence of exercise-induced angina indicated the likely existence of heart disease for both men and women. Moreover, potential high risk factors for women such as resting ECG being either normal or hyper and slope being flat were also identified. Only, resting ECG being hyper was shown to be a significant factor for men. The author had also stated that for women, resting ECG status is an important factor for heart disease prediction. Healthy status for both genders include slope being up, number of coloured vessels being zero, and oldpeak being less than or equal to 0.56.

Ilayaraja and Meyyappan [15] devised a method to predict the risk level of the patients having heart disease through frequent itemsets and the dataset of various heart disease patients had been used for this research work. Minimum support value and chosen symptoms had been used in the generation of frequent itemsets that would help the medical practitioner to make diagnostic decisions and determine the risk level of patients at an early stage. The authors

had also suggested that the proposed method can be applied to any medical dataset to predict the risk factors with risk level of the patients based on chosen factors. The authors had also shown that the risk level of patients can be identified efficiently from frequent itemsets.

Evanthia et al [17] proposed the state-of-the-art of machine learning methodologies applied for the assessment of heart failure. Particularly, models predicting the presence, estimating the subtype, assessing the severity of heart failure and predicting the presence of adverse events, such as destabilizations, re-hospitalizations, and mortality are presented. The authors investigated three machine learning algorithms namely support vector machine, neural network, classification and regression trees. The authors claimed that out of the three machine learning algorithms SVM outperforms better than that of rest of the algorithms.

Anooj [5] had presented a weighted fuzzy rule-based clinical decision support system (CDSS) for the diagnosis of heart disease, automatically obtaining knowledge from the patient's clinical data. CDSS for the risk prediction of heart patients consists of two phases: One being the automated approach for the generation of weighted fuzzy rules and the other was to develop a fuzzy rule-based decision support system. Mining techniques, attribute selection and attribute weightage method have been used in the first phase to obtain the weighted fuzzy rules. In the second phase, the fuzzy system was constructed in accordance with the weighted fuzzy rules and the attributes that were chosen. The author stated that his CDSS obtained better accuracy.

Rajeswari et al [6] proposed a method that made use of Artificial Neural Network for selecting the interesting or important features from the input layer of the network. A Multi Layer Perceptron Neural Network is used for selection of interesting features from an Ischemic Heart Disease (IHD) data base with 712 patients. Initially the number of attributes was 17 and after feature selection the number of attributes was reduced to 12. During training of the classifier the accuracy is 87 % and during testing phase of the classifier accuracy is 82%.

Akhil Jabbar et al. [11] proposed a new algorithm which combines KNN with genetic algorithm for effective classification of heart disease. The author also stated that Genetic algorithms perform global search in complex large and multimodal landscapes and provide optimal solution. From the authors of experimental results it has been observed that their proposed algorithm obtained better accuracy in diagnosis of heart disease.

Liu et al. [14] developed a novel scoring system for predicting cardiac arrest within 72 h and it was developed based on a semi-supervised learning algorithm namely manifold ranking, which had explored the local and global consistency of the data. From the extensive simulations results of the authors it can be inferred that manifold ranking based system attained better results than that of geometric distance scoring system in terms of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV).

Arabasadi et al. [18] proposed a highly accurate hybrid method for the diagnosis of coronary artery disease which had increased the performance of neural network by approximately 10% by enhancing its initial weights using genetic algorithm. The authors had claimed that they have achieved accuracy, sensitivity and specificity rates of 93.85%, 97% and 92% respectively, on Z-Alizadeh Sani dataset.

Yang and Garibaldi [13] proposed an information extraction system that was developed to automatically identify risk factors for heart disease in medical records. The authors had relied on quite a few nature language processing (NLP) techniques such as machine learning, rule-based methods, and dictionary-based keyword spotting. The proposed system had shown improved performance on the challenge test data with an overall micro-averaged F-measure of 0.915.

Long et al [16] proposed a heart disease diagnosis system using rough sets based attribute reduction and interval type-2 fuzzy logic system (IT2FLS) to handle with high-dimensional dataset challenge and uncertainties. A hybrid learning process comprising fuzzy c-mean clustering algorithm and parameters tuning by chaos firefly and genetic hybrid algorithms were employed in IT2FLS. To increase the performance of IT2FLS and to reduce the computational load, the rough sets based attribute reduction using chaos firefly algorithm had been examined to find optimal reduction. The author claimed that the experimental results had shown that the proposed system is better than other machine learning techniques namely Naive Bayes, support vector machines, and artificial neural network.

Alizadehsani et al [8] used a dataset called Z-Alizadeh Sani with 303 patients and 54 features and had also proposed a feature creation method to enrich the dataset. To determine the effectiveness of CAD, the author had used the features like Information Gain and confidence. In addition to the created features by means of Information Gain, Chest Pain, Region RWMA2, and age were determined to be the most effective ones. Experimental results had shown that the features like Q Wave and ST Elevation had the highest confidence and the accuracy rate was 94.08%

Tayefi et al. [19] had aimed to establish a predictive model for coronary heart disease using a decision tree algorithm. In this paper, the authors had used a dataset of 2346 individuals including 1159 healthy participants and 1187 participant who had undergone coronary angiography (405 participants with negative angiography and 782 participants with positive angiography). The authors had entered 10 variables of a total 12 variables into the decision tree algorithm (including age, sex, FBG, TG, hs-CRP, TC, HDL, LDL, SBP and DBP). It has been identified that the associated risk factors of CHD with sensitivity, specificity, accuracy of 96%, 87%, 94% and respectively.

Lee et al. in [1] and [2] proposed a unique methodology to develop the various features of heart rate variability (HRV) and carotid arterial wall thickness. They had also proposed a prediction model to improve the reliability of medical examinations and treatments for cardiovascular disease. Six classification methods were used to analyze the features and CPAR and SVM classification methods had been proved to be better. The authors also analyzed HRV for three recumbent postures. The interaction effects between the recumbent postures and groups of normal people and heart patients were observed based on HRV indexes. The authors measured intima-media of carotid arteries and used

measurements of arterial wall thickness as other features. Patients underwent carotid artery scanning using high-resolution ultrasound devised in a previous study. In order to extract various features the authors tested six classification methods. As a result, CPAR and SVM (gave about 85%-90% goodness of fit) outperforming the other classifiers. In their another work in [2], various experiments on linear and nonlinear features of HRV to evaluate classifiers were conducted and from their experiments, the authors claimed that SVM and Bayesian classifiers outperformed the other classifiers.

Narasimhan and Malathi [12] used fuzzy logic to classify the risks associated with coronary artery heart disease (CAHD) in female diabetic patients. The input parameters chosen by the authors were plasma glucose concentration, diastolic blood pressure, body mass index and age for designing Mamdani type fuzzy inference system. The risk of CAHD is predicted as low, intermediate and high. The PIMA women diabetes dataset had been used for simulation and it was implemented through MATLAB 2012a. Classification accuracy, sensitivity and specificity were used as performance metrics and the outputs were demonstrated using rule viewer and surface viewer. Also the authors further published a work on Fuzzy Logic System for Risk-Level Classification of Diabetic Nephropathy [20].

Paredes et al [9] made use of MATLAB tool to predict the risk of death/myocardial infarction for coronary artery disease (CAD) patients, within a short period of time. The authors had claimed that using this Matlab tool, the physician can calculate the CVD risk of a patient as well as perform a set of configurations to adjust the parameters of the global model to a specific population which can help in the decision support process. Two real patient datasets (Santa Cruz hospital 460 patients; Leiria-Pombal hospital centre 99 patients) were taken to test this tool against the other tools.

Kim et al [7] developed various classifiers to identify patients in high risk of CAC using statistical and machine learning methods, and compared them with performance accuracy. For statistical approaches, linear regression based classifier and logistic regression model were developed. For machine learning approaches, the authors suggested three kinds of ensemble-based classifiers (best, top-k, and voting method) to deal with imbalanced distribution of their data set. Their ensemble voting method outperformed all other methods including regression methods as AUC was 0.781.

Hartati [3] proposed a clinical decision support system (DSS) is a computer tool which uses two or more items of data to generate patient or encounter-specific advice. Kohenen ANN has been used as a model for the prediction of risk factor based coronary artery disease (CAD). Real life data had been used to train the system and in the proposed method 6 steps were used in the computation for training process. Their proposed model was successfully implemented and tested and the success rate was 89.47%.

Elbedwehy et al [4] proposed a computer-aided diagnosis system of the heart valve disease using binary particle swarm optimization and support vector machine, in conjunction with K-nearest neighbor and with leave-one-out cross-validation. Their system was applied in a representative heart dataset of 198 heart sound signals, which come both from healthy medical cases and from cases suffering from the four most usual heart valve diseases: aortic stenosis (AS), aortic regurgitation (AR), mitral stenosis (MS) and mitral regurgitation (MR). Their introduced approach starts with an algorithm based on binary particle swarm optimization to select the most weighted features which was followed by performing support vector machine to classify the heart signals into two outcome: healthy or having a heart valve disease, then its classified the having a heart valve disease into four outcomes: aortic stenosis (AS), aortic regurgitation (AR), mitral stenosis (MS) and mitral regurgitation (MR). From their experimental results the authors claimed that the overall accuracy offered by the employed approach is high compared with other techniques.

III. PERFORMANCE METRICS

True positive (TP): Sick people correctly identified as sick
False positive (FP): Healthy people incorrectly identified as sick
True negative (TN): Healthy people correctly identified as healthy
False negative (FN): Sick people incorrectly identified as healthy

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

$$sensitivity = \frac{TN}{TN + FP}$$

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

IV. FINDINGS AND CONCLUSIONS

This review article emphasizes several research works carried out for classification of heart diseases. The following are the findings of this survey research work.

- Decision support systems are only compatible for certain datasets and could not be used globally.
- Feature selection task plays an important role in classification of heart diseases.
- Machine learning algorithms can be employed in predictive data mining tasks.
- Conventional data mining algorithms could not achieve better accuracy when compared with soft computing techniques / machine learning algorithms.
- Bio-inspired particle swarm optimization algorithm helps in improving accuracy of the classifier.
- When compared with existing machine learning algorithm research works, SVM outperforms better.
- In most of the research works, time complexity performance metric has not been considered.

REFERENCES

[1] H. G. Lee, K. Y. Noh, H. K. Park and K. H. Ryu, "Predicting Coronary Artery Disease from Heart Rate Variability Using Classification and Statistical Analysis," 7th IEEE International Conference on Computer and Information Technology (CIT 2007), Aizu-Wakamatsu, Fukushima,2007,pp.59-64.

[2] H. G. Lee, K. Y. Noh and K. H. Ryu, "A Data Mining Approach for Coronary Heart Disease Prediction using HRV Features and Carotid Arterial Wall Thickness," 2008 International Conference on BioMedical Engineering and Informatics, Sanya, 2008, pp. 200-206.

[3] S. Hartati, "A Kohonen artificial neural network as a DSS model for predicting CAD," 2010 International Conference on Distributed Frameworks for Multimedia Applications, Yogyakarta, 2010, pp. 1-5.

[4] M. N. Elbedwehy, H. M. Zawbaa, N. Ghali and A. E. Hassanien, "Detection of heart disease using binary particle swarm optimization," 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), Wroclaw, 2012, pp. 177-182.

[5] P.K. Anooj, Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, Journal of King Saud University - Computer and Information Sciences, Volume 24, Issue 1, January 2012, Pages 27-40.

[6] K. Rajeswari, V. Vaithiyanathan, T.R. Neelakantan, Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks, Procedia Engineering, Volume 41, 2012, Pages 1818-1823.

[7] H. Y. Kim et al., "Identifying relatively high-risk group of coronary artery calcification based on progression rate: Statistical and machine learning methods," 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, 2012, pp.2202-2205.

[8] Roohallah Alizadehsani, Jafar Habibi, Mohammad Javad Hosseini, Hoda Mashayekhi, Reihane Boghrati, Asma Ghandeharioun, Behdad Bahadorian, Zahra Alizadeh Sani, A data mining approach for diagnosis of coronary artery disease, Computer Methods and Programs in Biomedicine, Volume 111, Issue 1, July 2013, Pages 52-61.

[9] S. Paredes, T. Rocha, P. de Carvalho, J. Henriques and J. Morais, "Matlab tool for cardiovascular disease risk prediction," 2013 2nd Experiment at International Conference (exp.at'13), Coimbra, 2013, pp.190-191.

[10] J. Nahar, T. Imam, K. S. Tickle, Y. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females," Expert Systems with Applications, vol. 40, pp. 1086 – 1093, 2013.

[11] M. Akhil jabbar, B.L. Deekshatulu, Priti Chandra, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, Procedia Technology, Volume 10, 2013, Pages 85-94.

[12] B. Narasimhan and A. Malathi, "A Fuzzy Logic System with Attribute Ranking Technique for Risk-Level Classification of CAHD in Female Diabetic Patients," 2014 International Conference on Intelligent Computing Applications, Coimbatore, 2014, pp. 179-183.

[13] Hui Yang, Jonathan M. Garibaldi, A hybrid model for automatic identification of risk factors for heart disease, Journal of Biomedical Informatics, Volume 58, Supplement, December 2015, Pages S171-S182.

[14] Tianchi Liu, Zhiping Lin, Marcus Eng Hock Ong, Zhi Xiong Koh, Pin Pin Pek, Yong Kiang Yeo, Beom-Seok Oh, Andrew Fu Wah Ho, Nan Liu, Manifold ranking based scoring system with its application to cardiac arrest prediction: A retrospective study in emergency department patients, Computers in Biology and Medicine, Volume 67, 1 December 2015, Pages 74-82.

[15] Ilayaraja M., Meyyappan T., Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets, Procedia Computer Science, Volume 70, 2015, Pages 586-592.

[16] Nguyen Cong Long, Phayung Meesad, Herwig Unger, A highly accurate firefly based algorithm for heart disease prediction, Expert Systems with Applications, Volume 42, Issue 21, 30 November 2015, Pages 8221-8231.

[17] Evanthia E. Tripoliti, Theofilos G. Papadopoulos, Georgia S. Karanasiou, Katerina K. Naka, Dimitrios I. Fotiadis, Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques, Computational and Structural Biotechnology Journal, Volume 15, 2017, Pages 26-47.

[18] Zeinab Arabasadi, Roohallah Alizadehsani, Mohamad Roshanzamir, Hossein Moosaei, Ali Asghar Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm, Computer Methods and Programs in Biomedicine, Volume 141, April 2017, Pages 19-26.

[19] Maryam Tayefi, Mohammad Tajfard, Sara Saffar, Parichehr Hanachi, Ali Reza Amirabadizadeh, Habibollah Esmaeily, Ali Taghipour, Gordon A. Ferns, Mohsen Moohebati, Majid Ghayour-Mobarhan, hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm, Computer Methods and Programs in Biomedicine, Volume 141, April 2017, Pages 105-109.

[20] B.Narasimhan, Dr. A. Malathi, "Fuzzy Logic System for Risk-Level Classification of Diabetic Nephropathy" in the IEEE International Conference on Green Computing, Communication and Electrical Engineering (ICGCCEE'14), 7th - 8th March, 2014.

