A Survey on Efficient Heart Disease Prediction Technique

Thilagamani S[a], and Anitha K[b]

a,b
Researc Department of Computer Science and Engineering, M.Kumarasamy College of Engineering, Thalavapalayam, Karur-639113, Tamilnadu, India.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: This paper is a survey on different heart disease prediction technique to predict and diagnosis of heart diseases. In order to increase the efficiency and accuracy of the prediction process an effective method is to be searched. For this searching process , data mining and machine learning techniques are recommended. The main purpose of this paper is to improve cardiac problem diagnostic method, by including early diagnosis warning method.

Keywords: Heart Disease, Machine Learning, Data Mining, early diagnostic warning system**.**

1. Introduction

Cardiac disease describes a variety of heart-related conditions. Heart conditions include a number of diseases, such as coronary artery disease, heart rhythm diseases (arrhythmias) and heart defects (congenital heart defects), among others. Heart disabilities include diseases of the blood vessel. The term cardiovascular disease is sometimes used interchangeably.Cardiovascular disease (CVD) usually refers to conditions involving the blood-vessels which are narrowed or blocked and which may cause heart attack (Myocardial infarction), stroke or chest pain.Other cardiac conditions such as those affecting the muscle, valves or rhythm of your heart are also considered as forms of heart disease. Each year, thousands of people death from CVDs, about 31% of the world's deaths.Today the health sector produces large amounts of patient information, diagnoses of diseases, etc. However, the researchers and practitioners do not use these data efficiently.Currently Quality of service is an important challenge for the healthcare industry . Quality of Service involves correct diagnosis and provides patients with effective treatment. Poor diagnosis can lead to serious and unacceptable consequences.There are different risk factors for heart disease. Family history, increasing age, ethnicity and malehood are some uncontrollable risk factors.

The process of data mining is the discovery of the unknown hidden patterns (knowledge) of large pre-existing data sets involving data mining, machine teaching, statistics and database systems.The knowledge discovered may be used to develop intelligent predictive decision-making systems for accurate diagnosis in different fields such as healthcare to offer affordable service and save precious life. Machine learning offers computer programmes the ability to learn from predetermined information and improve performance without interference by human beings and then use what they have learned to make an informed decision. Their performance will be improved by every successfull decision maker training programme.

The following figure shows the process of data discovery (KDD).

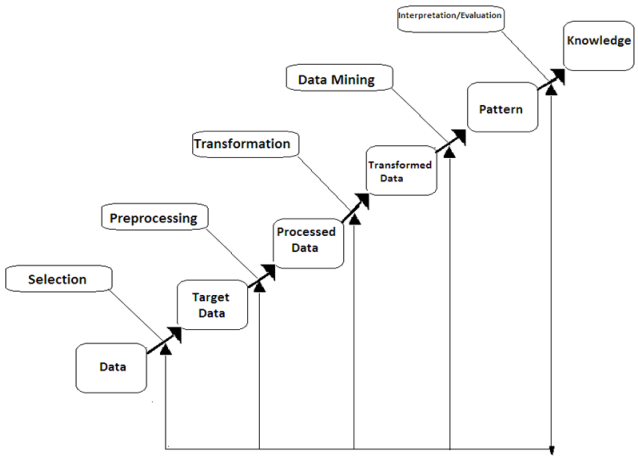


Figure 1 : Steps in Knowledge Discovery Process

2. Machine Learning Techniques

2.1 Classification:

Classification is a supervised data mining and software development technique . This is a two-tier process. The first step is called the learning step in which the model is built up and trained by a newly constructed dataset of class labells (training set), and the second step is the classification process (testing).

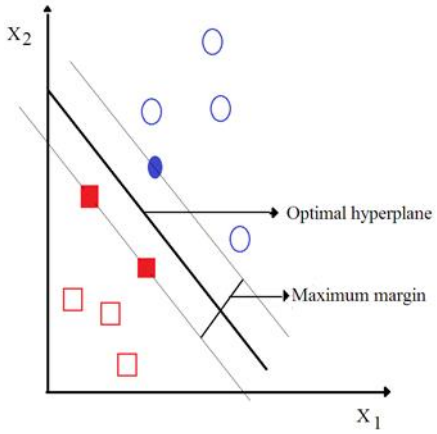
This paper critically summarizes different classification techniques

2.2 Decision Tree:

Decision tree is a technique used as a tool for supporting decision making using tree-like graphs or decision model. It takes a record or object described by a set of attributes as input and returns a "decision with a forecast input value."The input characteristics may be discreet or permanent. The decision tree comes after performing a sequence of tests.Each non-leaf node of a decision tree is the test for the corresponding attribute value, with the branches of the node indicating the possible test results.The value (decision) to be returned if that leaf is reached in every tree leaf node is specified. Multidimensional data can be handled. The Logistic Tree Model (LTM) and the Random Forest (RF) are algorithms for implementing the Decision Tree.

2.3 Support vector machine:

Support vector machine is the one of the supervised learning algorithm in machine learning. Regression, Classification and Outlier detection is the most powerful method in support vector machine. It is also called maximum margin classifier. SVM can handle both liner and non linear data. The training data is translated by a non linear transformation method into n-dimensional data. To seperate the transformed data it finds the best hyperplane. Hyper Plane separates the classes which are given.By maximising the hyper plane margin, classification is performed.



2.4 Random forest:

Random forests are supervised methods for the development of a classifying tree.In this input data vector algorithm, a new object from the input feature vectors can be classified in each tree of the forest .For each tree in the forest, the "votes" are considered to be classified for each tree and tree with the highest "vote." Choose "k" from the overall m characteristics randomly.Disperse the element into certain elements with the best division. Algorithms are able to accomplish the task of classification and regression. The strong and more accurate results are a collection, no decision-making trees, the more tree in the forest.

2.5 Naive Bayes:

A Naive Bayes Classification is a supervised Bayes theorem-based machine learning technique.As such, a Naive Bayes classification assumes that a certain class attribute is presence or lack independent of the presence or absence of any other class attribute.Naïve Bayes is used to calculate a post-class probability based on conditional probability to classify data sets. It is often used for the calculation and decisions about higher probabilities of later observations.

The equation is as follows:

$$P(A/B) = \frac{P(B/A)*P(A)}{P(B)}$$

Where B is the instance to be predicted, A is the instance class value.

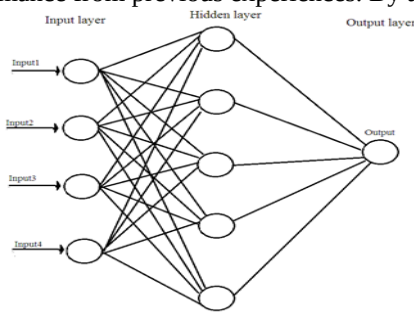
This formula is used to determine the class in which function that should be classified.

2.6 K-Nearest Neighbor(KNN):

KNN is a supervied classifier.It is more suitable for the classification algorithm.KNN finds the distance of nearest training data class labels in the presence of K value to predict the target label of a new test data point.KNN normally uses a K variable value of 0 to 10 to calculate the number of nearest exercising data points distance.For continuous variables the Euclidean distance function is used, and the Hamming distance function for categorical variables is applied.

2.7 Artificial Neural Networks:

Artificial neural networks are non-linear data processing capabilities of machine learning algorithms. Artificial neural network is a mathematical model or computational model based on a biological neural network sometimes . This is also a biological neural system emulation.Based on the internal or external features its structure can be changed. Neurak network contains Input and output layers, and a number of hidden layers are hidden between them.They are great tools for finding complex data patterns and continuously improving performance from previous experiences. By adjusting its weight and structure the error rate can be minimised.



2.8 Genetic Algorithm:

Genetic algorithm is an optimization solution based on a natural, biological development-friendly selection process.The genetic algorithm changes an individual solution population iteratively.Individuals are randomly chosen to produce children for the next generation as parent(s) from the current population through the genetic algorithm.The population is "changing" over the next few generations to an optimum solution. Chromosomes are represented in genetic algorithm solutions.Chromosomes consist of genes that represent the problem and are individual elements. The population is called the collection of all chromosomes. The genetic algorithm uses at each step the following 3 main kinds of rules (operators) from the population: a) Selection of individuals is used for reproduction purposes. b)Crossover is used to combine two parents into the next generation of children. C) In the search for a better solution, Mutation is used to change the new solutions.Mutation Prohibits GA at a local minimum to be trapped.

3. Related Work

3.1 Literature Survey

Ali Khazaee [1] proposed classifying heart beats with the application of particle swarm optimisation in addition to the definition of three types of electrocardiogram rhythms. The Extraction Function Module, the Classifier Module and the Optimization Module are three main modules. The Extraction Module function suggests an appropriate set incorporating the shape features and the timing features. In the classification module, a multi-class SVM-based classifier is proposed. For the optimization module, the best value is proposed for the SVM parameters using the particulate swarm optimization algorithm.The result indicates a very high degree of precision in the proposed algorithm.

Techniques:Support vector machine and Particle swarm optimization

Demerits:Limited in features

Future scope:Extend the features to classify the diseases

FidaBenish et.al.[2] focuses on the diagnosis of heart disease and suggests a classifier ensemble approach to strengthen the judgment of classifiers for the diagnosis of heart disease. The homogeneous collection is used for the classification of heart disease and ultimately the results are optimized with the use of the genetic algorithm. Comparison of this approach with the current Ensemble Technique revealed major increases in the precision of classification

Techniques:Genetic algorithm

Demerits:Difficult to predict the base classifiers

Future scope:Improve the fitness values

EI-Bialy R et.al in[3] also established a feature study of the collections of cardiac coronary artery disease data to expand the integration of machine learning outcomes with different CAD-related data collection. When the resulting trees are extracted and compared from different data sets, the fast decision tree and cut c4.5 tree are used. Popular aspects of these data sets are derived and used in every data collection in the later study for the same disorder. The results show that the precision of classifying the collected dataset is better than the average accuracy of classifying the multiple datasets.

Techniques:Decision tree construction

Demerits:Diversity problem can be occurred

Future scope:Learn the data problems

The key aim of Lee HeonGyu et.al in[4] is to classify coronary artery diseases by linear and nonlinear features of HRV and suggest a quantitative HRV test and an accurate predictive model to help enhance the efficiency of cardiovascular disease diagnostic evaluation.. HRV measures of the time and frequency domain as a linear

measure. Princare plots and uncertainty are used as nonlinear measures and then as a statistical tool. Several supervised learning approaches have been tested for estimation. The diagnosis of cardiovascular disease was checked by SVM and Associated Classifier (CMAR), Decision Tree Inference (C4.5), and Bayesian Classifiers. The findings also suggest that the diagnosis can be changed.

Techniques:Bayesian classifiers

Demerits:Supervised approach

Future scope:Implement unlabeled data

Singh Jagwant et.al, [5] focuses on refining the process of genetic algorithms and neural networks for classification of cardio-vascular diseases. GA and neural networks have their benefits paired with anticipation of the possibility of heart failure. In a variety of implementations, learning that reflects the optimal system behaviour is used in datasets. At a time when datasets contain learning about the structure to be compiled, the neural system ensures a response so it can be compiled from datasets.

Techniques:Neural network

Demerits:Local minima problem may be occurred

Future scope:Overcome the minima problems

In order to conduct a survey of established data exploration techniques in databases, JyotiSoniUjma Ansari et al [6] centred on the use of data mining technologies, which are in current physicians, particularly heart disease prediction. A series of research has been carried out to assess the efficiency of the data-mining predictive technique in a single collection of data and the results suggest that the Decision Tree is over performing and that the Bayesian classification has a near accuracy of a decision-making tree.

Techniques:KNN algorithm

Demerits:Time complexity can be high

Future scope:Analyze the time complexity

Sudhakar K in[7], which focuses on the study of cardiac disease prediction using data mining, aims to examine the various predictive data mining strategies proposed in recent years for Heart disorder diagnosis. Techniques for data processing can retrieve secret information from vast datasets. It helps to find the relationships and trends of the results. Data mining is used for a range of purposes, such as business companies, e-commerce and the healthcare, science and engineering sectors. Data mining is primarily used in the health care sector to forecast diseases from datasets. In this survey article, we researched and examined how data mining methods, such as sorting, clustering, fuzzy system and correlation laws, are used to forecast heart disease. This paper further points out the benefits and drawbacks of existing methods. It also addresses the future improvement of current works.

Techniques:Fuzzy system

Demerits:Need manual intervention

Future scope:Improve the optimization

Thenmozhi K et.al in [8] indicated that cardiac disease can be forecast using various tree decision-making strategies. The method of exploration of information is structured in multiple phases, while the first phase consists of compilation of data from different sources, the second phase of pre-processing of collected data, the third phase of data conversion into a format that is suitable for further processing and the fourth stage, if necessary, data mining. Methodology for data mining is applied to data for the recovery of usable understandings, and evaluation is the final step. This disease triggers complexities of life threats, such as heart failure and death. The importance of data mining is realised in the field of medicine and the necessary strategies in the field of disease prediction are applied. Although various classification techniques are widely used for Disease Prediction, for their simplicity and precision the Decision Tree classification is chosen. Selection tests for attributes such as information benefit, gain ratio, Gini index and calculation of distance can be used.

Techniques:Decision tree classifier

Demerits:Complexity is high

Future scope:Learn some effective techniques

A genetic algorithm based on a qualified repetitive fuzzy neural network in the diagnosis of heart diseases, KaanUyarAhmetIlhan [9] provides a genetic algorithm for the diagnosis of heart disease based on a trained repetitive neural fugacious network (RFNN). This study contains the Cleveland cardiac Disease University of California Irvine (UCI) data collection. 252 are used for preparation, 45 of which out of a total of 297 patient observations are selected. The results showed that the test sample was 97.78 percent correct. The root mean square error and the probability for error are calculated in addition to the accuracy, the specificity, the sensitivity, the precision and the f-score. On the basis of a comparison, the findings were satisfactory.

Techniques:Recurrent fuzzy neural networks (RFNN)

Demerits:Computational steps are high

Future scope:Need to analyze more datasets

LathaParthiban et.al[10] reflects on the topic of smart heart disease diagnosis. The approach is designed to provide reliably-compensated services and to eliminate medical errors, increase patient safety, avoid unnecessary medication changes and optimise patient outcomes through the inclusion of clinical decision support on computer-based health records. This Essay proposed a new approach to cardiovascular forecasting, based on the

CANFIS (Coactive neurofuzzy inference method).The CANFIS model is suggested with a perfect conceptual background approach, paired with a genetic algorithm to detect the presence of the disease. The model is adaptive, neural network functionality. CANFIS model performance has been tested in terms of consistency of the training performance and the results suggest that there is a good possibility for predicting heart attack in the proposed CANFIS model.

Techniques:Coactive neuro-fuzzy inference system (CANFIS)
Demerits:Difficult to analyze the various attributes
Future scope:Improve the algorithm to provide optimal results

Vasighi Mahdi et.al [11] implemented a new methodology to achieve strong predictive precision with a limited number of features. The suggested solution may be paired with different classifiers to enhance the efficiency of classification and the collection And notably discriminatory elements. Start with a variety of pre-selected filter features. In order to investigate the space of the function subsets we used GA, combined with Fisher's Linear Discrimination Analysis (LDA). Our suggested approach uses GA to assess the fitness of a single candidate subclass, and uses the LDA classification. In order to assess the utility of the final level function array, an external test set was chosen using Kohonen auto-ordering charts.We remember that in a small number of ways our approach will achieve high predictability.In patients without an angiographic procedure that may pose a high risk of mortality to people the recommended treatment would be used to diagnosis CHD.

Techniques:Genetic algorithm
Demerits:Outliers may be occurred
Future scope:Algorithm to overcome the risk in classification

NaharJ et.al in [12] centred on the classification comparison for cardiac disease diagnosis. The paper also revealed the effects of automatic role choice and a driven functional selection process based on medical expertise (MFS). The test results showed that MFS use greatly improves the drills, in particular in Conditions of precision, in Some of the data sets classifiers. MFS & MFS & CFS therefore offer Methods promising to be utilised in the diagnosis of Cardiac sickness. The individual risk factors are not especially susceptible to the risk of heart disease for everyone. The literature argues that for most patients with recognised unstable angina .This measure is not usually a sensible judgement (chest pain) or the person with an irregular ECG.

Techniques:Computerized feature selection process (CFS)
Demerits:Large number of features are extracted
Future scope:Analyze favorable strategy for heart disease data

Patil SB et.al in[13] focusing on eliminating large trends from heart disease storages for prediction. The identification of cardiac disease by many causes or symptoms is a dynamic problem which often has unforeseen consequences and is not free from erroneous hypotheses. This paper suggested an effective method for estimation of heart attack for the retrieval of important trends from cardiac warehouses. The data warehouse is originally prepared to suit the mining process.

Techniques:Frequent Pattern Mining
Demerits:Manual clustering can be constructed
Future scope:Learn automated approach

Chauhan Shraddha et.al in [14] based in an evaluation of its economic effect on the growing prevalence of cardiovascular diseases in India. Recent statistics indicate that CVDs outweigh caste, locality and economic barriers. It also puts an immense social pressure on creative citizens. Cardiovascular disorders (CVDs) are a trend towards increasing the reach of the younger age groups in the Indian population. The rise in CVD death and morbidity accounting is likely to lose India, avoiding a saving population change. Indian population. This reduces the competitiveness of the world in the otherwise favourable process of population change due to the economic cost of disease.

Techniques:CVD analysis
Demerits:Only handle limited datasets
Future scope:Improve the algorithm to predict more risk factors

Vanisree K et.al in[15], Proposed a form of decision support for a congenital heart disease condition that concentrates on the use of neural network signs and symptoms. With the implementation of the Back Propagation Neural Network the proposed framework was programmed and built using MATLAB's Interface feature. The Back Distributed Neural Network in this study is a multi-level Feed Forward Neural Network equipped by a supervised delta learning regime.

Techniques:Back-propagation Neural Network
Demerits:Manual diagnosis
Future scope:Need to improve multiple attributes

4. Performance Evaluation

The accuracy of various existing algorithm on heart disease prediction are listed below

ALGORITHM USED	ACCURACY
DECISION TREE	89.1%
SUPPORT VECTOR MACHINE	92.6%
K-NEAREST NEIGHBOR	90.9%
NAÏVE BAYES	89.1%

Table:Performance Evaluation

On analysing the literature survey some problems has been identified ,Only a labeled data based disease classification is occured. It also provides high number of false positive. Binary classification is occured. Computational complexity is high. To overcome these problems data mining techniques can be used to find some limitation for data analysis, like accuracy, speed, error rate, etc.

5. Conclusion:

To summarize, a comprehensive survey highlighting different classification techniques used for the heart disease prediction have been presented .Classification techniques has been analysed .Through deep learning algorithm heart disease prediction can be done in an effective way.The classification rule algorithm namely multi-layer preceptron can be used to improve the accuracy of the heart disease prediction with early diagnosis warning system.

References

1. Ali Khazaee. Heart beat classification using particle swarm optimization. Intell. System Applications. 2013:25–33.
2. FidaBenish, Nazir Muhammad, NaveedNawazish, AkramSheeraz. Heart disease classification ensemble optimization using genetic algorithm.IEEE; 2011. p. 19–25. in Computer and Communication Engineering,vol.3,Issue 5,May 2015.
3. EI-Bialy R, Salamay MA, Karam OH, Khalifa ME. Feature analysis of coronary artery heart disease data sets. ProcediaComput. Sci. 2015;65:459–68Computer Systems 28(2012).
4. Lee HeonGyu, Noh Ki Yong, RyuKeun Ho. Mining biosignal data: coronary artery disease diagnosis using linear and nonlinear features of HRV. LNAI 4819: emerging technologies in knowledge discovery and data mining. May 2007. p. 56–66.
5. Singh Jagwant, KaurRajinder. Cardio vascular disease classification ensemble optimization using genetic algorithm and neural network. Indian J. Sci. Technol. 2016;9(S1).
6. JyotiSoniUjma Ansari, Sharma Dipesh. Predictive data mining for medical diagnosis: an overview of heart disease prediction‖. Int. J. Comput. Appl. March 2011;17
7. JyotiSoniUjma Ansari, Sharma Dipesh. Predictive data mining for medical diagnosis: an overview of heart disease prediction‖. Int. J. Comput. Appl. March 2011;17(0975 – 8887).
8. Thenmozhi K, Deepika P. Heart disease prediction using classification with different decision tree techniques. Int J Eng Res Gen Sci 2014;2(6).
9. KaanUyarAhmetIlhan. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. 9th international conference on theory and application of soft computing, computing with words and perception. Budapest, Hungary: ICSCCW; 2017. 24-25 Aug 2017.
10. LathaParthiban, Subramanian R. Intelligent heart disease prediction system using CANFIS and genetic algorithm. Int. J. Biol. Biomed. Med. Sci. 2008;3(No. 3).

11. Vasighi Mahdi, Ali Zahraei, BagheriSaeed, VafaeimaneshJamshid. Diagnosis of coronary heart disease based on Hnmr spectra of human blood plasma using genetic algorithm-based feature selection. Wiley Online Library; 2013. p. 318–22.
12. Nahar J, Imam T, Tickle KS, Chen YPP. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert SystAppl 2013;40(1):96–104.
13. Patil SB, Kumaraswamy YS. Extraction of significant patterns from heart disease warehouses for heart attack prediction. Int. J. Comput. Sci. Netw. Secur(IJCSNS) 2009;9(2):228–35.
14. ChauhanShraddha, AeriBani T. The rising incidence of cardiovascular diseases in India: assessing its economic impact. J. Prev. Cardiol. 2015;4(4):735–40.
15. Vanisree K, JyothiSingaraju. Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. Int J ComputAppl April 2011;19(6). (0975 8887).