International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

# Prognosis of Cardiac Disease using Data Mining Techniques: A Comprehensive Survey

**D. Haripriya[1], Dr. M. Lovelin Ponn Felciah[2]**

[1]Research Scholar, [2]Assistant Professor

[1,2]Department of Computer Applications, Bishop Heber College, Trichy, Tamil Nadu, India

***How to cite this paper:*** D. Haripriya | Dr. M. Lovelin Ponn Felciah "Prognosis of Cardiac Disease using Data Mining Techniques: A Comprehensive Survey" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.1212-1216, **https://doi.org/10.31142/ijtsrd26605**

IJTSRD26605



**ABSTRACT**

The Healthcare exchange generally clinical diagnosis is ended commonly by doctor's knowledge and practice. Computer Aided Decision Support System plays a major task in the medical field. Data mining provides the methodology and technology to modify these rises of data into valuable data for decision making. By utilizing data mining techniques it requires less time for the prediction of the diseases with more accuracy. Among the expanding research on coronary diseases predicting system, it has happened significant to classifications the exploration results and gives readers with a layout of the current coronary diseases forecast strategies in every discussion. Data mining tools can respond to exchange addresses that expectedly being used much time over riding to decide. In this paper we study different papers in which at least one algorithm of data mining used for the prediction of coronary diseases. As of the study it is observed that Naïve Bayes Technique increase the accuracy of the coronary diseases prediction system. The commonly used techniques for Heart Disease Prediction and their complexities are outlined in this paper.

***KEYWORDS:*** *Data Mining, Cardiac diseases prediction, Naïve Bayes*

## 1. INTRODUCTION

Coronary disease is the kind of infection that includes the heart or veins. In today's world, huge number of population is experiencing different kinds of heart sicknesses more; the number of patients experiencing and dying this sickness is expanding day by day. So there is a need of exact and early detection of coronary illness with genuine and sufficient treatment which can save the life of numerous patients. However, because of the complicated procedures and different side effects furthermore, obsessive tests the right determination of heart maladies is a troublesome undertaking and causes delay in the appropriate treatment. Hence, there is a need to develop up the expectation frameworks for coronary illness which can help the medical specialists in the early and accurate analysis of coronary illness. In the techniques that we utilize the methods coordinated with the medicinal data framework then it would be more invaluable and it will diminish the expense too. This can be done in the wake of looking at different data mining techniques for finding their appropriateness. Data mining combines statistical analysis, machine learning algorithms and database technology for extracting the hidden patterns from the large databases. The coronary illness conclusion relies upon clinical and sullen information. The medical experts are helped by heart sickness expectation framework in anticipating the status of heart sickness and it is done dependent on the clinical information of patients. Classification and prediction are the most common used demonstrating objectives. Neural network, Naive Bayes, Genetic algorithm, Decision Tree, classification via clustering, Support Vector Machine (SVM) are some techniques used here.

## 2. TECHINQUES USED IN DATA MINING

To utilizing data mining techniques we can recognize illness at beginning time and we can totally cure the ailment by proper determination. Health care industry gather huge measure of data, which are not mined to find hidden data. Cure of this issue is data mining technique. Data mining is the way toward investigating large arrangement of data and summarizing into valuable data. The below table shows the different Data mining Tasks and Intelligent Techniques and are [1]

1. Classification
2. Clustering
3. Association

Table1: Data Mining Tasks and Intelligent Techniques

| S. N. | Data Mining Task | Data Mining Algorithm & Technique |
|---|---|---|
| 1 | Classification | Decision Trees, Rule-based, Neural Networks, Naive Byes and Bayesian Belief Networks, Support Vector Machines, Genetic Algorithms |
| 2 | Clustering | K-Means |
| 3 | Regression and Prediction | Support Vector Machines, Decision Trees, Rule induction, NN |
| 4 | Association and Link Analysis (finding correlation between items in a dataset) | Association Rules Mining (ARM) |
| 5 | Summarization | Multivariate Visualization |

## 1. Classification

Classification is one of the familiar problems in data mining. To classify the data/objects into different classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics etc.

### 1.1. Decision tree

There are numerous decision tree algorithms, among them the most well known is J48 which utilizes a pruning system to construct a decent decision tree. Pruning is a strategy which attempts to take out the over fitting data which isn't so pertinent in settling on a decision and prompts poor prediction. At last, a tree is worked to give adaptability and exactness balance. The decision tree approach is all the more powerful for classification issues. There are two stages in this systems building a tree and applying the tree to the dataset. There are numerous well known choice tree calculations CART, ID3, C4.5,CHAID, and J48. From these J48 algorithm is utilized for this framework. J48 calculation utilizes pruning technique to manufacture a tree. Pruning is a strategy that reduces size of tree by leaving over fitting data, which prompts poor exactness in predications. The J48 algorithm recursively classifies data until it has been ordered as perfectly as could be allowed. This strategy gives most extreme exactness on preparing information. The general idea is to manufacture a tree that gives parity of adaptability and precision.

### 1.2. Naïve Bayes

It is a basic system for building classifiers, modals that assign the class labels to the issue occasions and represented to as vectors of highlight esteems where class labels are drawn from limited set. Naives Bayes is anything but a solitary calculation for preparing such classifiers yet it is a group of calculations dependent on some normal rule All Naïve Bayes classifiers expect that the estimation of specific component is autonomous of the estimation of some other element when class variable is given.

## 2. Clustering

Clustering is a data mining techniques that influences important or useful cluster of objects that to have comparative characteristic utilizing automatic technique. Not quite the same as classification, clustering system likewise characterizes the classes and place questions in them, while in classification objects are appointed into predefined classes. For example In prediction of heart diseases by utilizing clustering we get group or we can say that list of patients which have same risk factor. Means this makes the different list of patients with high glucose and related risk factor and so on. *1*

### 2.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbor (KNN), a supervised learning model too, is utilized to classify the test data utilizing the tests straightforwardly. In KNN, an object is classified by the larger part of its nearest neighbors. On the other hand, the class of a new sample is Predicted dependent on some distance measurements where the distance metric can be a basic Euclidean distance. In the working advances, KNN first calculates k (No. of the nearest neighbors). From that point forward, it finds the distance between the training data and afterward sorts the distance. In this manner, a class label will be assigned to the test data dependent on the majority voting.

### 2.2. K-Medoid Algorithm

K- Medoid algorithms is used to discovering Medoid in a cluster which is center position points in a cluster. The simple scheme of a K-Mediod cluster algorithm is to find ki clusters in N objects by first randomly finding a representative object for every cluster. The every parallel object are clustered with the Medoid. It uses the representative objects are reference points rather than of taking the mean value of the elements in every cluster. The algorithm precedes the input factor of ki, and the number of clusters to be partitioned into a set of N objects. Thus, K-Medoid is more robust as compared to K-Means [2].

## 3. Association Rule:

Association rule mining is a very important rule of data mining techniques. Association rule is distinguishing of association huge data base and their qualities.

## 4. Neural Network

Neural Network is a parallel, distributed data handling structure comprising of various amounts of preparing components called nodes, they are interconnected by means of unidirectional signal channels called connections. Each preparing component has a single output that branches into numerous connections and each passes on the equivalent signal. The NN can be arranged in two main groups as indicated by the manner in which they learn. They are supervised learning and unsupervised learning.

In supervised learning the network compute a reaction to each input and after that contrasts it and the objective esteem. If the compute chance that the registered reaction varies from the objective esteem, the loads of the system are adjusted by a learning rule. Instances of directed learning are Single-layer perceptron and Multi-layer perceptron. In unsupervised learning the systems learn by recognizing extraordinary highlights in the issues they are exposed to. Example for unsupervised learning is self-organized feature maps.

## 3. DATA MINING TOOLS

There are various data mining tools used for data mining purpose. These are WEKA, TANAGRA, MATLAB and .NET FRAMEWORK. [3]

**WEKA:** It is a data mining tool which was developed in New Zealand by the University of Waikato that implements data mining algorithms using JAVA language. WEKA is acollection of machine learning algorithms and their application to the data mining problems. These algorithms are directly applied to the dataset. WEKA supports data file in ARFF format. [3]

**TANAGRA:** It is open source software as researchers can access to the source code and add their own algorithms and compare their performances, if it conforms to the software distribution license. [3]

**MATLAB:** It is a data mining tool built in high level language. It provides interactive environment for visualization, numerical computation and programming. The built in math functions, language and tool explore various approaches and helps to reach a solution faster than with the spreadsheet of traditional programming languages like C,C++ and JAVA. It analyse data, develop algorithms, and create models and applications. [3]

**NET FRAMEWORK:** It is a software framework developed by Microsoft which runs primarily on Microsoft windows. It provides secure communication and consistent applications. It provides language interoperability (each language can code written in other languages) across several programming languages. [3]

## 4. LITERATURE SURVEY

H. Benjamin Fredrick David et al. [4] proposed in this paper, the UCI data repository is used to compare the three calculations, for example, Random Forest, Decision trees and Naive Bayes. From this paper,it has been experimentally proved that Random Forest gives ideal outcomes as compare with Decision tree and Naive Bayes. The Future work of this paper can be had to create an effect in the exactness of the Decision Tree and Bayesian Classification for extra improvement subsequent to applying genetic algorithm so as to diminish the real information for procuring the ideal subset of characteristic that is sufficient for coronary illness expectation. The automation of coronary illness forecast utilizing real continuous information from health care organizations which can be constructed utilizing big data. They can be sustained as a streaming data and by utilizing the data, examination of the patients continuously can be prepared.

Md. Fazle Rabbi et al. [5] designed this paper as coronary illness is one of the crucial causes to death, it should be to be accurately identified at all around beginning time to get recovery from it. Now and again, genuine expert will most likely be unable to recognize the sickness because of some absence of Knowledge and proper experiences. In this way, computer based capability precise expectation framework might be a choice to distinguish the coronary illness for fixing it right away. Thus, in this paper, three for the most part utilized information mining order methods, for example, SVM, KNN and ANN have been examined and assessed utilizing standard Cleveland coronary illness dataset. It has been dissected that RBF portion based SVM can beat KNN and ANN based on the order rate while KNN is likewiseoffering preferred execution over ANN. This similar investigation likewise prescribes that the essentially assessed classifier can be utilized for ongoing expectation of coronary illness patients and for anticipating the hazard factor of heart disappointment with the end goal of guaranteeing extra consideration so beginning period heart disappointment can be stayed away from. Be that as it may, all the more preparing information whether from emergency clinics or from space specialists can be included for expanding the forecast execution of the classifiers. In addition, assorted component decrease methodologies may likewise be connected on the dataset for getting improved execution.The main objective of our work is to provide a study of different data mining techniques that can be used in automated heart disease prediction systems. Various data mining classifiers are defined in this work which has emerged in recent year for effective and efficient heart disease diagnosis. The analysis shows that different technologies are used in all the papers by using different number of attributes. So different technologies used show different accuracy to each other. In some papers it is shown that SVM provide effective and efficient accuracy about 85% as compared to other data mining techniques in prediction of heart disease. So applying data mining techniques help heath care professionals in the diagnosis of heart disease is having success, the use of data mining techniques to identify a suitable treatment for the heart disease patients has received less attention.

Megha Shahi et al. [6] proposed this paper is to give an investigation of various data mining methods that can be utilized in computerized coronary illness expectation frameworks. Different data mining classifiers are characterized in this work which has developed in late year for effective and efficient coronary illness conclusion. The examination demonstrates that diverse advances are utilized in every one of the papers by utilizing distinctive number of traits. So unique advances utilized show distinctive exactness to one another. In certain papers it is demonstrated that SVM give successful and effective precision about 85% when contrasted with other information mining systems in forecast of coronary illness. So applying information mining strategies help heath care experts in the determination of coronary illness is having achievement, the utilization of information mining procedures to recognize an appropriate treatment for the coronary illness patients has gotten less consideration.

Kanika Pahwa et al. [7] proposed this paper of examination is to classify the data in two classes either in positive or in negative outcome for coronary illness. A hybrid approach of feature selection is received to upgrade the classification problem, consolidated result of SVM-RFE and, gain-proportion are utilized to get subset of feature and remove irrelevant or redundant feature. On subset of highlights Naïve Bayes and Random forest are connected to characterize them into presence or absence of disease.It has been appeared in results that precision improved for the two classifiers when connected to selected features. Proposed approach of feature selection not just diminished size of dataset yet in addition upgraded the performance of both the classifiers models.

Ritika Chadha et al. [8] proposed this paper prediction system by using ANN, Decision Tree and Naive Bayes methods. They executed it by using C# and also used the Python platform. According to this research paper, the prediction rate or accuracy for each of the data mining technique was calculated. Based on the observations or technical experiments, it was found that Artificial Neural Networks gave highest accuracy surveyed by Decision Tree and Naive Bayes respectively. The accuracies of each of the

technique are as follows: ANN achieved an accuracy of 90%, Decision tree got accuracy of 88.02% and accuracy of 85.86% was obtained by the Naïve Bayes algorithm.

Jagdeep Singh et al. [9] designed this paper different association and classification strategies are executed on the heart datasets to anticipate the heart infections. The association algorithm like Aprior and FPGrowth are utilized to discovers affiliation guidelines of heart dataset attributes. Classification algorithms are utilized to predict small set of relationships between credits in the databases to manufacture an exact classifier. The proposed cross hybrid associative classification is executed on weka condition. The relative outcomes demonstrate that IBk (k Nearest Neighbor) with Aprior cooperative calculations creates preferred outcomes over others. At long last a specialist framework is produced for the end client to check the danger of heart sicknesses based on given parameters and the best cooperative arrangement procedure. The exploratory outcomes demonstrate that huge number of the guidelines support in the better find of heart sicknesses that even help the heart specialist in their conclusion decisions.

Marjia Sultana et al. [10] proposed this paper tends to the issue of prediction of heart disease as indicated by some input attributes. The coronary illness turns into a plague all through the world. It can't be effectively predicted as it is a difficult task that requests mastery and higher learning for prediction. Data mining removes covered up data that assumes a critical job in settling on choice. This paper played out a trial utilizing diverse data mining methods to discover a progressively precise procedure for the heart infection forecast. In this paper, two data sets(gathered and UCI standard) are utilized independently for every data mining method.Our findings show that for coronary illness prediction performance of Bayes Net and SMa classifiers are the ideal among the examined five classifiers: Bayes Net,SMa, KStar, MLP and J48.

K.Gomathi Kamaraj et al. [11] proposed this paper we have examine some of effective techniques that can be utilized for heart illnesses classification and the accuracy of classification techniques is assessed dependent on the selected classifier algorithm. A vital test in data mining and machine learning zones is to fabricate exact and computationally effective classifiers for Medical applications. The execution of Naive Bayes indicates abnormal state contrast and different classifiers.

Ilayaraja M et al. [12] have developed a strategy to create frequent item sets dependent on the client's clinical data (symptoms). The discoveries helped them to gauge the hazard degree of patients influenced. Frequent item sets were delivered dependent on the selected symptoms and minimum support value. The gained frequent item sets helped the specialists to make diagonist ends and helped them to know the likelihood of dangers in patients at a beginning time. The strategy can be connected to any medicinal dataset to predict the probability of dangers with hazard dimension of the sufferers dependent on chose factors. The study demonstrated that the created technique could discover the likelihood dimension of patients proficiently from successive thing sets. Other than this they have contrasted the execution of this technique and strategies like apriori, semi-apriori and association rule mining algorithm dependent on example generation.

Shahed Anzarus Sabab et al. [13] proposed this paper we tried to concentrate on the significance of feature selection in cardiovascular ailment prognosis treatment utilizing diverse data mining algorithms. Using proper attribute selection strategy, any order calculation can be improved fundamentally. Attribute with less commitment in dataset, frequently miss lead the classification model and results in poor prediction accuracy. In our work, we found that Naïve Bayes gave best outcome before attribute selection But after performing out a controlled and careful feature selection, SVM ended up being the best classifier Area under ROC curve analysis indicated results to support us where every one of the three classifier demonstrated much better upgrades after feature selection method, In addition to this work we will endeavor to assess some more up to date calculations with better feature selection techniques.

Table2: Diagnosis Of Heart Diseases Used Different Data Mining Techniques:

| S. No | Author | Techniques | Accuracy |
|---|---|---|---|
| 1 | H. Benjamin Fredrick David et al.(2018) | Random forest | 80% |
| 2 | Md. Fazle Rabbi et al.(2018) | Support Vector Machine(SVM),<br>K-Nearest Neighbour(KNN),<br>Artificial Neural Network(ANN) | 85%<br>80%<br>73% |
| 3 | Megha Shahi et al.(2017) | SVM(Support Vector Machine) | 85% |
| 4 | Kanika Pahwa et al.(2017) | Naive Bayes,<br>Random Forest | 83%<br>82% |
| 5 | Ritika Chadha et al.(2016) | Artificial Neural Networks(ANN),<br>Naive Bayes | 85% ,<br>88% |
| 6 | Jagdeep Singh et al.(2016) | Naive Bayes | 97% |
| 7 | Marjia Sultana et al.(2016) | J48 | 86% |
| 8 | K.Gomathi Kamaraj et al.(2016) | Naïve Bayes, Artificial Neural Networks(ANN),<br>J48 | 79%<br>76%<br>77% |
| 9 | Ilayaraja M et al.(2015) | Association Rule Mining Algorithms | 85% |
| 10 | Shahed Anzarus Sabab et al.(2015) | Naive Bayes | 86% |

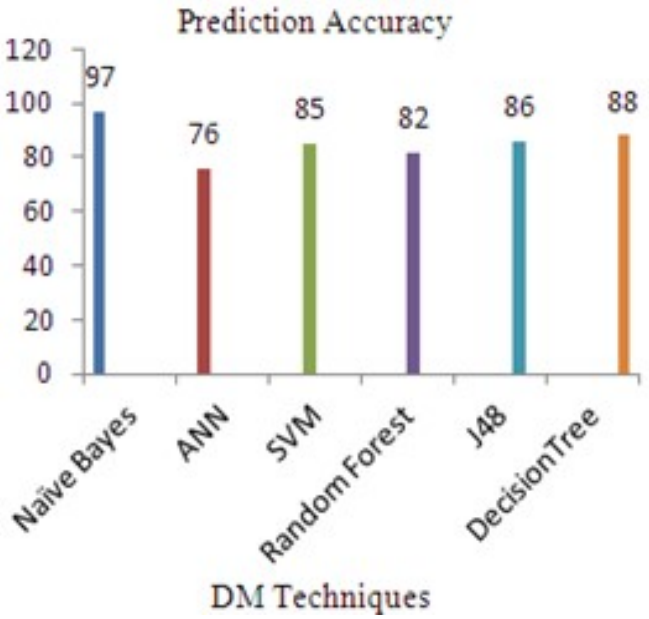


Figure.1: Graphical representation of the data mining techniques reviewed.

## 5. CONCLUSION

On observing the various data mining techniques for the prediction of cardiac diseases, the Naïve Bayes shows more accuracy than other techniques. The following table has been examined from the extensive study on the various algorithms in the prediction of cardiac diseases. This review provides recommendations for the researchers and cardiologists to cooperate to perform simple clinical datasets for the data mining models.

## REFERENCES

[1] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review", International Journal of Computer Applications (0975 – 8887) Volume 136 – No.2, February 2016.

[2] Arora, P., Deepali, Varshney, S., "Analysis of KMeans and K-Medoid Algorithm for Big Data", International Conference on Information Security and Privacy (ICISP2015), Volume.78, 2016.

[3] Era Singh Kajal, Nishika, "Prediction of Heart Disease using Data Mining Techniques", International Journal of Advance Research, Ideas and Innovations in Technology. © 2016, IJARIIT All Rights Reserved Page | 1 ISSN: 2454-132X (Volume2, Issue3) Available online at: www.Ijariit.com.

[4] H. Benjamin Fredrick David, S. Antony Belcy," Heart Disease Prediction Using Data Mining Techniques", ISSN: 2229-6956 (Online) Ictact Journal On Soft Computing, October 2018, Volume: 09, Issue: 01 Doi: 10.21917/Ijsc.2018.025.

[5] Md. Fazle Rabbi, Md. Palash Uddin, Md. Arshad Ali, Md. Faruk Kibria, Masud Ibn Afjal1, Md. Safiqul Islam and Adiba Mahjabin Nitu," Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction", American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN : 2320-0936 Volume-7, Issue-2, pp-278-283 www.ajer.org.

[6] Megha Shahi, Er. Rupinder Kaur Gurm" Heart Disease Prediction System Using Data Mining Techniques- A Review" International Journal Of Technology And Computing (IJTC) ISSN-2455-099X, Volume 3, Issue 4 April 2017.

[7] Kanika Pahwa, Ravinder Kumar, "Prediction of Heart Disease Using Hybrid Technique For Selecting Features", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)GLA University, Mathura, Oct 26-28, 2017.

[8] Ritika Chadha, Shubhankar Mayank, "Prediction of heart disease using data mining techniques" ,Springer , December2016.

[9] Jagdeep Singh, Amit Kamra, Harbhag Singh," Prediction of Heart Diseases Using Associative Classification",978-1-5090-0893-3/16/$31.00 ©2016 IEEE

[10] Marjia Sultana, Afrin Haider, Mohammad ShorifUddin" Analysis of Data Mining Techniques for Heart Disease Prediction" 978-1-5090-2906-8/16/$31.00 ©2016 IEEE.

[11] K. Gomathi Kamaraj, D. Shanmuga Priyaa" Heart Disease Prediction Using Data Mining Classification", www.ijraset.com Volume 4 Issue II, February 2016 IC Value: 13.98 ISSN: 2321-9653.

[12] Ilayaraja M, Meyyappan T , " Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Item sets", Elsevier 2016.

[13] Shahed Anzarus Sabab, Ahmed Iqbal Pritom, Md. Ahadur Rahman Munshi, Shihabuzzaman," Cardiovascular Disease Prognosis Using Effective Classification and Feature Selection Technique.