Data Mining Techniques in Prediction of Coronary Artery Disease: A Survey

[1]S. Omprakash, [2] Dr. M. Ravichandran
[1]Research Scholar, [2] Associate Professor
[1,2]Department of Computer Science,
[1,2] SRMV College of Arts and Science , Tamil Nadu, India.

Abstract : The usage of data mining is getting evolved in an unbelievable speed, due its applicability in most streams. The size of the data in this current world is getting increased day by day. Due to the enormous usage of web based application data are being captured lively. There exists no use of simply storing the data in computer machines. It's necessary to use mining concept to find the hidden data which is undetected before. Data mining has started stepping in medical field to identify and predict the diseases. This paper surveys the data mining methods and recent proposals towards classifying the coronary artery disease, which is getting increased among the peoples in the world. Also, this paper attempts to provide different solutions in identifying and predicting the coronary artery disease.

IndexTerms **- Coronary artery disease, Classification, Prediction.**

I. CORONARY ARTERY DISEASE (CAD)

Cardiovascular diseases are arising as a top reasons major death in the world. CAD is one of the type of cardiovascular disease. More than 25% of the people affected by CAD may die unexpectedly without having any symptoms. CAD is considered as a worst disease which affects the heart leading to heart attacks. The different heart diseases are congestial cardiovascular defects, rheumatic fever disease, arteries disease, high blood pressure, congestive failure, and CAD. Among the different cardiovascular disease, CAD is main reason towards death, where the percentages crosses 53. The heart in the human body is a physically powerful muscular pump. It is conscientious to move nearly 3000 gallons of body every day. The human heart is in the need of constant blood supply to function well. Human heart needs proper supply of blood, if not means then the heart may become weak due to lack of oxygen and other nutrients to function in a good manner. This leads to pain in the chest, namely angina.

The main reason for the death in the developed countries is CAD. Even though it got passed more than a decade, still the system to detect and treat the CAD gets lack. CAD becomes a root reason for death in India. CAD has become a multifaceted disease with the reason of atherosclerosis. CAD is a developing disease among people due to hereditary, environment, and change of life style. This disease evolves mainly from family history, where all people know that CAD can be medicated but in-time diagnosis is important. It's necessary to analyze the history of CAD affected family, where it is considered as globally important. Changes in the characteristics and habits of family, leads a major way for CAD. Family aspects of the CAD are partially explained by medical associations and experts. The other factors such as obesity, smoking, hypertension, dyslipidemia and diabetes are also a reason for CAD. The parents having the myocardial infarction (MI), increases the risk of CAD double times. The risk level gets increased if both father and mother are affected by MI.

Also, history of the family is a important factor of risk in atherosclerosis. Sometimes family history are not fully considered for identifying a risk among the individual person. When the heart did not get the sufficient blood, some symptoms are found in the body, which include burning sensation, pain in the chest, squeezing, tightness, unexpected level of sweating, short breathing,

II. DATA MINING

Data mining is the progress of finding significant information from the available data that is stored and being stored in the repositories. It is done by utilizing the technology of pattern recognition, mathematical and statistical methods. It can also be said as; Data mining is the process of observing the dataset in order to find the relationship that is not identified before, and performing the summarization which will be helpful for the data owner. Data mining is not for a specific field like computer science, it can be applied to different fields like marketing, engineering, education and even in medical field. In medical field, data mining is used to classify the patients affected by disease and also it is used to predict the diseases for the persons.

A. Knowledge Discovery from Data (KDD)

The current world treats data mining as KDD process. The KDD involves the below mentioned steps.

- **Cleaning of data:** removal of noise and inconsistent data from the available data.
- **Integration of data:** data gathered from multiple sources are combined.
- **Selection of data:** relevant data is retrieved from database.

- **Transformation of data**: data are transformed and converted into a specified format.
- **Mining of data:** better methods are applied in the transformed data to extract the patterns.
- **Evaluation of pattern:** to find the true patterns denoting the knowledge measure.
- **Presentation of knowledge:** making the evaluated pattern of data to visualize for better understanding.

B. Classification

It is the procedure of identifying a strategy which is used to describe and differentiate the data concepts. The strategies that are derived were based on the analysis made on training data. Different strategies were used to identify the class of an object. The benchmark strategies include neural network, regression, and relevance analysis.

- **Neural Network:** It is a collection of neuron-like processing units, where each units is linked with the weighted connection between different units. Some of the neural network based classification models are k-nearest neighbor, support vector machine, Bayesian classification.
- **Regression:** It is a statistical based method which is mostly used for prediction. It includes the process of identifying the distribution trends based on the data available.
- **Relevance:** It is a model of making an attempt to find attributes which are closely related to classification. The attributes will be selected for the classification and regression process based on demand.

C. Classification Vs Prediction

Here is the criteria for comparing the methods of Classification and Prediction −

- **Accuracy –** Classification accuracy denotes the classifier's ability. It can perform the prediction with the label of classes more correctly, where classification accuracy of predictor denotes how well a specified predictor be able to guess the assessment of predicted attribute for a original data.
- **Speed** − This denotes the cost of the computation towards producing and usage of the classifier.
- **Robustness –** It denotes the classifiers ability to create right predictions from the noisy data.
- **Scalability –** It denotes classifiers ability to construct or predict more effectively even when a huge amount of data is given.
- **Interpretability** − It denotes the level of the classifier towards understanding the extent.

III. RECENT PROPOSALS

A. Optimal Feature Selection Method

This method [1] was proposed to filter the features in the heart disease dataset and diagnosing heart disease, where modified differential evolution algorithm was used to perform feature selection for cardiovascular disease and optimization of selected features. The feature selection gets weak due to using the differential algorithm concept, results in low accuracy.

B. Hybrid Reasoning Based Method

This method [2] was proposed to predict the CAD. It used combinatorial advantage of Fuzzy set theory, k-nearest neighbor and case-based reasoning helps to yield enhanced prediction results. Disease Prediction Support System has facilitated the healthcare services, data security and privacy are still crucial challenging issues to be addressed. Sensitivity results very low and Specificity results very high, which is unexpected.

C. Feature Identification Based Method

This method [3] aimed to identify significant features and data mining techniques that can improve the accuracy of predicting cardiovascular disease, where prediction models were developed using different combination of features and classification techniques.

D. Validation Based Prediction Model

A Cox proportional hazards regression model was used to develop risk prediction model [4] for cardiovascular diseases. The risk assessment ability of the developed model was evaluated, and a bootstrapping method was used for internal validation. The predicted risk was translated into a simplified scoring system. A decision curve analysis was used to evaluate clinical usefulness.

E. Attribute Reduction Method

Heart disease diagnosis system [5] using rough sets based attribute reduction and interval type-2 fuzzy logic system. It utilizes a hybrid learning process comprising fuzzy c-mean clustering algorithm and parameters tuning by chaos firefly and genetic hybrid algorithms.

F. Ensemble Method

The combination [6] of Fuzzy concept, k-nearest neighbor and case-based were utilized to predict heart disease. It extends the prediction using patients' sensitive information and it was evaluated with the statistical evaluation metrics, where the classification accuracy got down.

G. Artificial Neural Method

Artificial Neural Cell System for classification [7] was proposed to predict the heart disease which was inspired by mechanisms that develop the brain and empowering it with capabilities such as information processing/storage and recall, decision making and initiating actions on external environment.

H. Hidden Markov Model Based Method

Hidden Markov Model based prediction method [8] was proposed to predict disease candidate genes for CAD through gene expression profiles. In this method, the disease genes were aimed to predict in order to find the CAD evolving from hereditary. The results came with increased false positives.

I. Optimization Based Method

Optimization based heart disease prediction system [9] was proposed by utilizing rough sets based attribute reduction and interval type-2 fuzzy logic system. The integration between rough sets based attribute reduction was aimed to handle with high-dimensional dataset challenge and uncertainties. The results came majority contents in true positive and true negative, which is unbelievable.

J. Nearest Search Method

An updated version of K-Nearest Neighbor algorithm and K-Sorting and Searching (KSS) were combined to frame a algorithm namely Nearest Search algorithm [10] for the prediction of CAD. The curve fitting mathematics method was utilized to improve the results. The results become weak when the dataset size gets increased.

IV. Z-ALIZADEHSANI DATASET

The dataset [17] consists of 303 patients records. Each record holds 54 features. Every feature in the dataset are considered as the indicators of CAD, which is according to medical literature, but some of the specific features are never been used in data mining approaches for CAD diagnosis. The features are arranged in four groups: demographic, symptom and examination, ECG, and laboratory and echo features. Each patient could be in two possible categories CAD or Normal. A patient is categorized as CAD, if his/her diameter narrowing is greater than or equal to 50%, and otherwise as Normal. Some of the features are HTN identifies the history of hypertension, DM is the history of Diabetes Mellitus, Current Smoker is current consumption of cigarettes, Ex-Smoker is the history of previous consumption of cigarettes, and FH is the history of heart disease in first-degree relatives.

V. PERFORMANCE MEASURES

In data mining, the performance of algorithms is measured using accuracy, sensitivity, and specificity. It is considered as most important due to its applicability in the field of medicine. The confusion matrix is a type of table which allows visualization of the performance of an algorithm. Considering two class problem (with C1 and C2 classes), the matrix will have two rows and two columns that specifies the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN).

- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$
- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $False\ Positive\ Rate = \frac{FP}{FP+TN}$
- $True\ Positive\ Rate = \frac{TP}{TP+FN}$

VI. MOTIVATION TOWARDS RESEARCH WORK

Considering future risk complication towards health, CAD results as a major problem among prediction of diseases. CAD is one of the major types of disease, where 25% of affected people die suddenly without any previous symptoms. Heart attacks become severe in patients who are all affected by CAD. It is estimated that CAD tends to affect the younger population and could negatively affect the productivity and workforce. World Health Organization estimates 11.1 million deaths all over the world from CAD in 2020. Being aware of CAD symptoms can aid in timely treatment, and reduce the severity of a disease's side effects. Cardiovascular diseases are becoming the most common reasons of death across South Asia countries. Correct and in-time diagnosis of CAD is very important. Many high-level treatments (i.e., Angiography) are available for CAD, but it has many side effects and it is costly. Performance degradation exists due to more number of records.

VII. PROBLEM STATEMENT

The healthcare industry has volumes data and that need to be mined to discover hidden information for effective decision making. Data mining algorithms are available to analyze the dataset and predict the results towards disease, but the performance and results gets varied for datasets. Most data mining algorithms are available for constant or specific dataset. When applying the

algorithms in different datasets, the performance of the algorithm gets weak which leads to a major problem in medical field. Time taken for analyzing and predicting the results are higher.

VIII. OBJECTIVE

The main objective of this research work is to propose data mining algorithms in order to make prediction by making classifying the patients affected by CAD. This research work has planned to solve the problems discussed in section 7 by proposing three different approaches (discussed in section 9).

IX. EXPECTED PROPOSED WORK TO ACHIEVE OBJECTIVE

A. Ant Colony Optimization based Support Vector Machine (ACO-SVM)

SVM is a traditional algorithm used to perform the classification and it is used widely. The classification accuracy provided by the SVM is not sufficient in medical field, just to increase the classification accuracy, ant colony based optimization will be utilized.

B. Principal Component Analysis based Support Vector Machine (PCA-SVM)

It is a well known fact that PCA and SVM perform the data mining operations in a great manner, but still it is not explored the concept of utilizing PCA and SVM together to classify and predict the CAD. PCA-SVM research work will be aimed to perform classification by reducing the dimensionality of the data.

C. Hidden Markov Model based Support Vector Machine (HVV-SVM)

HMM is a tool based on statistics utilized for modeling generative sequences characterized by a set of observable sequences. HMM-SVM aims to classify and predict CAD with increased accuracy, where HMM is used to find the hidden information towards CAD that is not observed by SVM.

X. CONCLUSION

This paper has started with the introduction of data mining and followed up with coronary artery disease. Data mining methods towards available for classification of coronary artery diseases were surveyed. The dataset available for classification of coronary artery disease, namely Z-Alizadehsani dataset was discussed. Metrics used to measure the performance of the data mining algorithms are discussed with this formula. Finally the paper concludes by discussing the motivation to proceed the research work with the strong objective. Further, the paper defined the works to be carried in future to meet the objective.

REFERENCES:

1. T. Vivekanandan, N. C. S. N. Iyengar, Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, Computers in Biology and Medicine, Volume 90, 2017, Pages 125-136.
2. D. Malathi, R. Logesh, V. Subramaniyaswamy, V. Vijayakumar, A. K. Sangaiah, Hybrid Reasoning-based Privacy-Aware Disease Prediction Support System, Computers & Electrical Engineering, Volume 73, 2019, Pages 114-127.
3. M. S. Amin, Y. K. Chiam, K. D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics, Volume 36, 2019, Pages 82-93.
4. T. Honda, D. Yoshida, J. Hata, Y. Hirakawa, Y. Ishida, M. Shibata, S. Sakata, T. Kitazono, T. Ninomiya, Development and validation of modified risk prediction models for cardiovascular disease and its subtypes: The Hisayama Study, Atherosclerosis, Volume 279, 2018, Pages 38-44.
5. N. C. Long, P. Meesad, H. Unger, A highly accurate firefly based algorithm for heart disease prediction, Expert Systems with Applications, Volume 42, Issue 21, 2015, Pages 8221-8231.
6. D. Tay, C. L. Poh, R. I. Kitney, A novel neural-inspired learning algorithm with application to clinical risk prediction, Journal of Biomedical Informatics, Volume 54, 2015, Pages 305-314.
7. O. Nikdelfaz, S. Jalili, Disease genes prediction by HMM based PU-learning using gene expression profiles, Journal of Biomedical Informatics, Volume 81, 2018, Pages 102-111.
8. N. C. Long, P. Meesad, H. Unger, A highly accurate firefly based algorithm for heart disease prediction, Expert Systems with Applications, Volume 42, Issue 21, 2015, Pages 8221-8231.
9. A. J. Amutha, R. Padmajavalli, D. Prabhakar, A novel approach for the prediction of treadmill test in cardiology using data mining algorithms implemented as a mobile application, Indian Heart Journal, Volume 70, Issue 4, 2018, Pages 511-518.