PREDICT THE RISK OF CARDIO VASCULAR DISEASE USING DATA MINING TECHNIQUES: A SURVEY

S.Tamil Fathima, K. Fathima Bibi[2]

[1]*Research Scholar in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli*
E-mail: tamilfathima@jmc.edu
[2]*Assistant Professor in Computer Science, Thanthai Periyar Govt. Arts & Science College (A), Tiruchirappalli*
E-mail: kfatima72@gmail.com

Abstract. *Heart disease is a very fast growing common disease and a major cause of death worldwide. Health care industry is considered as one the most information intensive industries. There is vital knowledge in the health care system and data mining technologies are commonly applied to enrich the data process. Data mining helps to make and predict the status of disease using health care data. Early detection of symptoms of heart disease is a serious challenge in the present situation. Research is being done to diagnose with hybrids of data mining techniques. The focus of this paper is to review the classification and data mining techniques which used for heart disease prediction.*

Keywords: *Heart Disease, Data Mining , Prediction, Feature Selection.*

1. INTRODUCTION

Meaningful patterns and associations of knowledgeable data can be discovered through data analysis using the process of data mining. Various methods and algorithms are applied in the data mining to extract useful knowledge in the form of pattern in the data. Knowledge Discovery Database (KDD) is considered as another name of Data mining. There are various steps involved in data mining such as Data Integration, Data Selection, Data Cleaning, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation and Decisions or use of Discovered Knowledge. Data mining divided into two models such as Descriptive Model and Predictive Model. Predictive Model is used to predict unknown or future values of other variables. Classification, Regression and Time-Series Analysis belong to the category of Predictive Model.

Descriptive Model is developed to issue a good knowledge of data without trying to attack a particular situation. Clustering, Association Rules, Sequence Discovery and Summarization belong to the category of Descriptive Model [1,2]. Data mining techniques such as clustering and classification methods such as Naive Bayes, Decision tree, Random forest and K-nearest neighbour are used to identify several heart diseases[3,4].

Medical Data mining is an activity of different effort which includes much inaccuracy and uncertainty [5]. Health care data is very big. The Medical Decision Support System was proposed to optimize medical errors and costs assist in earlier disease detection and to achieve preventive medicine. Some incidents will end in mistakes, huge medical cost will demage the level of support and help to the patients[6].The data mining has been used in medical domain to make medical decisions and diagnose medical problems.

2. HEART DISEASE

Heart Disease is also referred as Cardio Vascular Disease(CVD).Heart disease is one of the major risks in which the whole word is fighting on predicting the risk in the earlier stage. The human heart is one of the main components of the body, circulating blood throughout the body through the blood supply system [7].If affects the brain and heart stops working completely and life loss happens within a few minutes[8].Heart attack is caused by abrasions of the heart muscle due to insufficient oxygen supply and inattention to pump blood[9].It is difficult to accurately diagnose true heart disease because the data are so complex[10]. some of the risk factors of the Cardio vascular disease. Smoking often causes heart attack. Various factors and risk of heart disease are chest pain, heart burn and stomach pain, pain in the arms and seating. A close and updated study done in 2018 by WHO reveals the outcome that 56.9 million life loss happened in the world during the year 2016 was only by heart disease[11].The most timely tests and efficient methods for heart disease are very important.

3. FEATURE SELECTION

There are different data sets to capture heart disease, but there are some features or attributes limited no of rarely used to diagnose the disease. Some data sets have redundant features. Negative effects are caused by unwanted features. They increase the training time and so many features having redundant and irrelevant features can be very inconvenient [12,13].Feature selection is a method of removing attributes with small or no information[14].Feature selection is an efficient way to removes the unnecessary, redundant and irrelevant features from the dataset. The feature only contains relevant and useful attribute elements so it helps to reduce the training time and improve the classification accuracy[15].

4. DATA SETS

The heart disease prediction used data set taken from UCI(University of californio, Irvine) data mining repository[16]. Data set is the collection of similar data records[17]. The data set used a number of records such as1)Cleveland:303 2) Hungarian: 294 3) Switzerland: 123

4) Long Beach VA: 200. All data sets contain 76 attributes but used only 14 attributes. This 14 attributes contain eight categorical attributes and six numerical attributes. Cleveland data set and statlog data set are very popular and commonly used[18]. Because Cleveland data set and statlog data set have minimum number of missing values but other data set have more number of missing values [17].

TABLE I: Heart disease dataset Attributes used

S.NO	Attributes	Description
1	Age	Age in years
2	Sex	Gender instance 1=male, 0=female
3	cp	Chest Pain type (1= typical angina,2=atypical, 3=non- angina pain, 4=asymptomatic)
4	Trestbps	Resting blood suger (in mm Hg on admission to hospital)
5	chol	Serum cholesterol in mg/d1
6	Fbs	Fasting blood sugar>120 mg/d1(1=true, 0=false)
7	restecg	Resting electrocardiographic results(0= normal, 1= having ST-T wave abnormality, 2=left ventricular hypertrophy)
8	thalach	Maximum heart rate
9	exang	Exercise Induced Angina (1=yes, 0=no)
10	oldpeak	ST Depression include by Exercise Relative to Rest
11	slope	Slope of the Peak Exercise ST Segment(1 = Up Sloping 2 = Flat 3 = Down Sloping)
12	ca	Number of Major Vessels Colored by Fluoroscopy
13	thal	Defect types 3 = Normal 6 = Fixed Defect 7 = Reversible Defect
14	num	Diagnosis of heart disease Class (0 = Healthy 1 = Have Heart Disease)

5. DATA MINING TECHNIQUES USED IN HEART DISEASE PREDICTION

Karthikeyan G et. al [19] proposed a hybridization method which combines linear stacking model and Xgboost algorithm (HLS-Xgboost) to improve prediction accuracy. This HLS-Xgboost model performs better than other models such as Decision Tree (DT), Naive Bayes (NB) classifier, Density base Spatial clustering model, SVM, Random Forest (RF), Multi-Layer perceptron (MLP), and Linear Regression (LR) respectively. It gives 96% more of accuracy than other existing models.

Vicky Singh et. al [20] proposed a recommendation system using machine learning algorithms based on vital data like cholesterol level, age, etc. for heart disease prediction. The proposed model gives 90% of accuracy and performs better than SVC and Decision tree classifier.

Md. Nahiduzzaman et. al [21] proposed two classifiers such as Multi-Layer Perceptron neural network (MLP) and another is Support Vector Machine (SVM). In addition, the heart disease is classified into two-class and five-class level in the proposed work. SVM, MLP gives 92.45%, 90.57% accuracy respectively in two-class classification problem. SVM, MLP produces accuracy of 59.01%, 68.86% in five-class classification problem. They concluded that SVM outperforms in two-class classification and MLP Performs better in five-class classification for heart disease diagnosis.

Devansh Shah et. al [22] presented a model which depends on supervised learning algorithms. K- nearest neighbor algorithm gives 90.7% of more accurate result than other supervised learning algorithms such as Naïve Bayes, decision tree, random forest.

Saba Bashir et. al [23] described the heart disease prediction using feature selection techniques and algorithms to enhance accuracy. Logistic regression SVM, Decision Tree, Naïve Bayes and Random forest are used in Rapid miner tool on a UCI dataset. The experimental result shows that the accuracy of Decision Tree, Regression, Random Forest, Naïve Bayes and Logistic Regression SVM of 82.22%, 82.56%, 84.17%, 84.24% and 84.85% respectively. They suggested
that Logistic Regression (SVM) is a better feature selection technique for heart disease prediction.

.Kanika Pahwa et. al [24] proposed a heart disease prediction model using hybrid feature selection SVM-RFE (support Vector Machine - Recursive Feature Elimination) algorithm for feature selection. UCI dataset been used in the proposed model. In addition, Naive Bayes, Random Forest were used as classifiers for categorize the disease existence or absence.

C. Sowmiya et. al [25] proposed a hybrid model for heart disease prediction using ACO (Ant Colony Optimization) with HKNN (Hybrid K-Nearest Neighbor) classifier. This ACO-HKNN is compared with existing KNN (K-Nearest Neighbor), C4.5, Naïve Bayes, Decision Tree and Support Vector Machine (SVM) classification techniques. The proposed hybrid prediction model produced a better accuracy of 99.2% when compared with other existing classification techniques.

Pooja Rani et. al [26] proposed a hybrid verdict carry system by combining the GA (Genetic Algorithm) and recursive feature elimination for feature selection. Synthetic Minority Oversampling Technique and standard scalar technique used for pre-processing. SVM, naive bayes, random forest, logistic regression and adaboost classifiers are used in the proposed system. It provides 86.6%, of accuracy.

Anchana Khemphila et. al [27] proposed an improved approach for Heart disease Classification using MLP (Multi Layer Perceptron) with BPLA (Back-Propagation Learning

Algorithm) and Feature Selection Algorithm (FLA). Attributes were reduced from Thirteen to eight and the accuracy differences is 1.1% in training data set and 0.82% in the validation data set.

Namariq Ayad Saeed et. al [28] proposed a Heart Disease Prediction System by combining BPSO (Binary Particle Swarm Optimization Algorithm) with Mutual Information (MI) filter. Logistic regression is one of the most important technique used for classification. This MI_BPSO produces a Classification accuracy of 98.33% when compared with BPSO. Execution time of the MI_PBSO been significantly improved when compared with existing BPSO.

Luxmi Verma et. al [29] proposed a hybrid model using CFS (Correlation based Feature Subset) with PSO (Particle Swam Optimization) and Kmeans clustering algorithms for diagnosis of Coronary artery disease (CAD). The proposed hybrid model provides 90.82% accuracy than other existing techniques such as MLR (Multinomial Logistic Regression), FURIA (Fuzzy Unordered Rule Induction Algorithm), MLP (Multi-Layer Perceptron) and C4.5.

Youness Khourdifi et. al [30] proposed the FCBF (Fast Correlation-Based Feature) selection to enhance the heart disease classification worth. In addition, the existing classification algoritms such as KNN, SVM, Multilayer Perception, Artificial Neural Network, NB, RF are optimized by PSO (Particle Swarm Optimization) combined with ACO (Ant Colony Optimization). This hybrid optimized model gives the 99.65% of classification accuracy.

J. Vijayashree et. al [31] presented a novel function based on population diversity and tuning function for identifying optimal weights. In addition, they proposed a fitness function for PSO with SVM in order to reduce attributes count and the accuracy improvement. The proposed PSO- SVM produces better accuracy than other existing feature selection algorithms.lications domain, strength and weakness of their algorithms.

M. Anbarasi et. al [32] proposed the GA with three classifiers like NB, classification by clustering and DT used to predict the diagnosis of patients with reduced number of features. Thirteen attributes are reduced to six attributes by using GA. The NB and classification by clustering having inconsistencies and high missing value but DT data mining techniques accu-racy is high 99.2% and less missing value.

Durgadevi Velusamy et. al [33] proposed a machine learning algorithm for effective diagnosis and prediction of CAD (Co-ronary Artery Disease). It contains a heterogeneous ensemble method which combines the classifiers such as Random For-est, KNN and SVM for effective diagnosis and ensemble vot-ing technique for CAD prediction. In feature selection, the Boruta wrapper based feature selection algorithm and SVM have been used based on attribute importance and rank. In ensemble voting technique, among based Majority-Voting (MVEn), Average Voting (AVEn), and Weighted-Average Voting (WAVEn), the WAVEn algorithm gives 98.97% of classification accuracy, 100% of sensitivity, 96.3% of speci-ficity, and 98.3% of precision for the original dataset. In the balanced dataset, the WAVEn algorithm achieves 100% of accuracy, sensitivity, precision and specificity in CAD diagnosis.

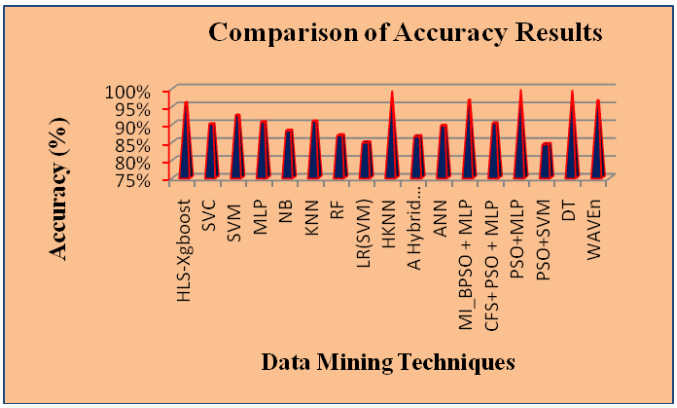
Jyoti Soni et. al [34] are focused on a detailed survey related to heart disease prediction using data mining techniques. In same dataset, Decision Tree, Bayesian Classification algorithm are performed better than other prediction models such as Neural Networks, KNN, Classification based on clustering. In addition, the accuracy of Bayesian classification, Decision Tree algoritm have been further improved by combining Genetic Algorithm. The comparison of all the techniques is summarized as given in Table 2.

Table 2: The accuracy accrued by the existing data mining techniques used for the prediction of CVD was summarized in the given below.

Author	Data Mining Techniques	Accuracy	Feature Selection
Karthikeyan et. al	HLS-Xgboost	96%	13
Vicky Singh et. al	SVC	90%	13
Md.Nahiduzzaman et. al	SVM MLP	92.45% 90.57%	13
Devansh et. al	NB KNN DT RF	88.157% 90.789% 80.263% 86.84%	13
Saba Bashir et. al	MRMR/FS LR(SVM)	84.85%	13
Kanika Ravinderkumar et. al	NB RF	84.1584% 84.164%	10 12
Sowmiya.C et. al	HKNN Hybrid KNN	99.02%	9
Pooja Rani et. al	A Hybrid combining GA& RFE	86.60%	8
Anchana Khemphila et. al	ANN	Training 89.56%	8
Namariq Ayad saeed et. al	MI_BPSO + MLP	96.66%	8
Luxmi Verma et. al	CFS+ PSO + MLP	90.28%	7
Youness Khourdifi et. al	PSO+MLP	99.65%	7
Vijayashree et. al	PSO+SVM	84.36%	6
M.Anbarasi et. al	DT	99.2%	6
Durgadevi Velusamy et. al	WAVEn	96.55%	5

6. RESULTS AND DISCUSSIONS

The comparative analysis among the existing data mining techniques HLS-Xgboost, SVC, SVM, MLP, NB, KNN, DT, RF, LR(SVM), HKNN, A Hybrid combining GA& RFE, ANN, MI_BPSO + MLP, CFS+ PSO + MLP, PSO+MLP, PSO+SVM, WAVEn is carried out. The performance of the existing data mining techniques are compared with respect to the accuracy results is as shown in the Fig.1.



7. CONCLUSION

Knowing the initial stage of the heart disease with data mining techniques is one of the biggest challenges in the health care sector. In health care sector different processes generate data in huge quantities. Some of the computerized health care monitoring systems and various medical instruments are continuously collecting health care data and hence the volume of the clinical data is increasing in rapid manner. Early detection of symptoms of heart disease can save people from this disease. The important task of this paper is to review the classification and feature selection of data mining techniques used for heart disease prediction. The techniques used in reviews produced good outcomes. This will be beneficial to the patients with heart disease if the accuracy is achieved by hybriding the available techniques.

8. REFERENCES

[1] Venkatadri.M. Dr.Lokanathan C.Reddy (2011), A Review on Data Mining from Past to Future, International journal of Computer Applications, Vol 15(7), pp.19-22.

[2] Leventhal, Barry (2010), An introduction to data mining and other techniques for advanced analytics. Journal of Direct, Data and Digital Marketing Practice 12, no. 2 (2010): 137-153.

[3] Rafiah Awang and Palaniappan.S (2007), Web based Heart Disease Decision Support System Using Data Mining Classification Modeling Techniques, Proeeding of iiWAS, pp.177-187.

[4] Patel J, TejalUpadhyay D, Patel S (2015), Heart disease prediction using machine learning and data mining technique. Heart Disease. Vol 7(1),pp.129–37.

[5] Monali Dey,Siddharth Swarup Rautaray (2014), Study and Analysis of Data Mining Algorithms for Healthcare Desicion Support System, International Journal of Computer Science and Information Technology, Vol 5(1), pp.470-477.

[6] Siri Krishnan Wasan, Vasutha Bhatnagar and Harleen Kaur(2006), The Impact of Data Mining techniques on medical diagnostics, Data Science Journal, Vol 5(19), pp 119-126.

[7] Tanya Lewis. (2016, March 22). Human Heart: Anatomy, Function & Facts [Online]. Available: https://www.livescience.com/34655 human-heart.html.

[8] Keerthana, T.K(2017),Heart disease prediction system using data mining method,Int. J. Eng. Trends Technol. Vol 47(6), pp 361–363.

[9] Sudha.A, Gayathri.P and Jaishankar.N(2012), Utilization of Data MiningvApproaches for prediction of life Threatening Disease Survivability, IJAC(0975-8887).

[10] Alia, A.F, Tawee (2017), A Feature selection based on hybrid binary cuckoo search and rough set theory in classification for nominal datasets. Inf. Technol. Comput. Sci. 4(April), 63–72.

[11] World Health Organization. The world health report 2000: health systems: improving performance. World Health Organization, 2000.

[12] Mohammad Ashraf Ottom, Girija Chetty, Dat Tran, and Dharmendra Sharma(2012), Hybrid approach for diagnosing thyroid, hepatitis, and breast cancer based on correlation based
feature selection and nave bayes. volume 7666, pages 272– 280.

[13] Dharmendra Modha and W. Spangler(2003). Feature weighting in k-means clustering. Machine Learning, Vol 52:217–237.

[14] Latha Parthiban and R.Subramanian(2007), Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International journal of Biological and Life Sciences, Vol 3(3), pp.157-160.

[15] TaoWang, ZhenxingQin,ZhiJin and ShichaoZhang(2010). Handling over-fitting in test cost-sensitive decision tree learning by feature selection, smoothing and pruning. Journal of Systems and Software, 83:1137–1147.

[16] Kurgan, Lukasz A., Cios, Krzysztof J., Tadeusiewicz, Ryszard, Ogiela, Marek, & Doodenday, Lucy S. (2001), Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine, 149-169.

[17] "UCI Machine Learning repository:", http://archive.ics.uci.edu/ml/datasets/ Heart Disease.
[18] Israa Nadheer1, Mohammad Ayache2, Hussein Kanaan(2021), Heart Disease Prediction System Using Machine Learning Algorithm, Journal of Advanced Computer Science & Technology, Volume 8, pp.23-31.
[19] Karthikeyan G, Komarasamy G, Daniel Madan Raja S(2021), An Efficient Method for Heart Disease Prediction Using Hybrid Classifier Model in Machine Learning, Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pp. 5708 – 5717.
[20] Vicky Singh, Brijesh Pandey(2021), Prediction of Cardiac Arrest and Recommending Lifestyle Changes to Prevent It using Machine Learning, International Conference on Intelligent Technologies & Science, pp. 1-6.
[21] Md. Nahiduzzaman, Md. Julker Nayeem, Md. Toukir Ahmed(2019), Prediction of Heart Disease Using Multi-Layer Perceptron Neural Network and Support Vector Machine, 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, pp 1-6.
[22] Devansh Shah, Samir Patel, Santosh Kumar Bharti(2020), Heart Disease Prediction using Machine Learning Techniques, SN Computer Science, pp. 1-6.
[23] Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, Khurram Bashir(2019),Improving Heart Disease Prediction Using Feature Selection Approaches, Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, pp. 619-623.
[24] Kanika Pahwa, Ravinder Kumar(2017), Prediction of Heart Disease Using Hybrid Technique For Selecting Features, 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University, Mathura, pp. 500-504.
[25] C. Sowmiya, P. Sumitra(2020), "A hybrid approach for mortality prediction for heart patients using ACO- HKNN", Journal of Ambient Intelligence and Humanized Computing, pp. 1-8.
[26] Pooja Rani, Rajneesh Kumar, Nada M. O. Sid Ahmed, Anurag Jain(2021), A decision support system for heart disease prediction based upon machine learning, Journal of Reliable
Intelligent Environments, Springer, pp. 1-13.
[27] Anchana Khemphila, Veera Boonjing(2011), Heart disease Classification using Neural Network and Feature Selection, 21st International Conference on Systems Engineering, IEEE Computer Society, pp. 406-409.
[28] Anchana Khemphila, Veera Boonjing(2011), Heart disease Classification using Neural Network and Feature Selection, 21st International Conference on Systems Engineering, IEEE Computer Society, pp. 406-409.
[29] Luxmi Verma, Sangeet Srivastava(2016),P. C. Negi, A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data, J Med Syst, vol 40, pp. 1-7.
[30] Youness Khourdifi, Mohamed Bahaj(2019), Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization, International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019, pp. 242-252,
http://www.inass.org/10.22266/ijies2019.0228.24
[31] J. Vijayashree, H. Parveen Sultana(2018), A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier, Programming and Computer Software, 2018, Vol. 44, No. 6, pp. 388–397, https://doi.org/10.1134/S0361768818060129
[32] M. Anbarasi , N.ch.s.n.iyengar ,n.ch.s.n.iyengar(2010), Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm , International Journal of Engineering Science and Technology Vol. 2(10), ISSN: 0975-5462, pp. 5370-5376.
[33] Durgadevi Velusamy, Karthikeyan Ramasamy(2021), Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset, Computer Methods and Programs in Biomedicine, Elsevier, Vol. 198, 2021, pp. 1-13, https://doi.org/10.1016/j.cmpb.2020.105770
[34] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni(2011), Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887), Vol. 17– No.8, pp. 43-48.