JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Survey on Smart Health Prediction Using Data Mining

[1] K L Sujitha, Assistant Professor, Department of Computer Science and Engineering, MVJ College of Engineering, India
Devatha Shivshyl, Chagi Vijetha, Chithra NB [234], Student, Department of Computer Science and Engineering, MVJ College of Engineering, India

Abstract : Imminent need of turning huge amount of available health data into useful information and knowledge attracts data mining techniques in medical diagnosis process. Data mining is a procedure of distinguishing and extracting valuable data and setting up connection between attributes in substantial datasets. Existing health disease prediction models use one or multiple data mining techniques. This paper surveys heart disease prediction systems systematically where in techniques are compiled, tabulated and analyzed based on hybrid techniques categorization. In this paper, the techniques are classified into two main categories: Discrete and Integrated, which are further classified as supervised, unsupervised, hybrid and miscellaneous. It is revealed from this survey, even though usage of one data mining technique performs well, hybrid data mining techniques yield promising outcomes in the determination of coronary illness.
Key words: Data mining, Health disease prediction (HDP).

1.INTRODUCTION

The most common and dangerous diseases in the world. As indicated by the Centre for Disease Control (CDC), coronary illness is the main source of death causing 1 in every 4 deaths in the U.S.A. is from health disease [28]. A heart disease prediction system can help physicians and patients to recognize health disease in early stages so as to prevent excessive damage.

TABLE 1. EXISTING SURVEY ON HEART DISEASE PREDICTION

Paper Citation	Claim/ Focus
[1]	The authors claim, different technologies generate different accuracy which also depends on data size, number of attributes considered and tools used for implementation.
[2]	The authors claim that the goal of this paper is to comprehend different information mining strategies that are accessible to anticipate the coronary illness.
[3]	The main objective is to give a detailed explanation of existing data mining algorithms. The authors concluded that decision tree algorithms have less error rate over Association rule and Bayesian classifier.
[4]	Patient data is used to find patterns and features using data mining with data analytics techniques to predict heart disease well before the consultation with medical practitioners. They attempted to reduce the datasets required for prediction with increase in accuracy.

A reliable model for predicting the existence of heart disease with known and unknown risk factors is required. High blood pressure, high cholesterol, etc., are the early symptoms of heart disease that are useful to predict heart diseases using data mining techniques. There exist few survey articles in the literature, which is given in Table I along with the main focus of the survey work.

Many research articles on heart disease prediction use multiple data mining techniques. Each technique has different corollary on accuracy of heart disease prediction. This paper discusses prediction techniques based on the number of attributes used for prediction

along with a comparative analysis to understand the state-of-art in this area. The number of attributes considered for prediction varies. Most of the papers used Cleveland Health disease dataset having 14 attributes as given in the Table II.

TABLE II. ATTRIBUTES IN CLEVELAND HEART DISEASE DATASET

Attribute	Description
Age	Age in Years
Sex	0 for Female and 1 for Male
cp	Value for typical Angin -1, atypical Angina -2 , nonanginal pain -3, Asymptomatic-4
Trestbps	Testing blood pressure
Chol	Serum cholesterol in mg/dl
Fbs	If the fasting blood sugar > 120 mg/dl, Fbs =1 else Fbs=0
Restecg	Results of Resting electrocardiographic which can be 0, 1 or 3 where o indicates normal result, 1 indicates presence of ST-T wave abnormality and 3 indicates probable or definite left ventricular hypertrophy.
Thalach	Max heart rate achieved
Oldpeak	ST depression induced by exercise with respect to rest
Exang	Whether exercise induced angina present then 1 else 0
Slope	Unstopping -1 , flat -2 and down sloping -3
ca	Number of significant vessels (0-3) hued by fluoroscopy
Thal	Normal-3; fixed defect -6; reversible defect- 7
num (Heart Disease)	No-0 Yes- 1

Supervised Discrete Data mining Techniques in HDP

There are four techniques - Naive Bayes, Decision Tree, Neural Network and others - used by authors from supervised learning for HDP as shown in Table 1.

Naive Bayes: A weight is assigned to each attribute based on the attribute having high impact on disease prediction and evaluating the performance [1]. A decision support is developed in the HDP system using Naive Bayes

Classification Technique using extracted hidden information from a historical health disease database [5]. The authors of paper [6] have used the technique of Naïve Bayes Classifier. The paper doesn't give clarity on dataset used for prediction. They have considered 8 attributes from the dataset. The authors claim that their technique gave better accuracy over the heart disease prediction have neither given accuracy measures nor compared with already existing heart disease prediction techniques.

Decision Tree Based: The authors of paper [7] discuss technology of Decision Trees (DT) based Rule Induction to provide a better, integrated and accurate prediction of heart disease. Dataset from the 5th Korean Health and Nutrition Examinations Survey (KNHANES V-1, 2020) having 17 attributes is used in this paper. The Decision Tree based Rule Induction consists of four basic steps: Decision Tree Creation; Pruning; Feasibility Evaluation; and Model Analysis and Prediction. The paper does not give clarity on all components considered in the proposed architecture. The authors claim that proposed prediction model is relied upon to add to the Korea heart disease prediction with better and exact outcomes.

Neural Network Based: multilayer perception and learning vector quantization algorithms are used as neural network model. The authors of paper [8] have proposed a new approach by executing the Multilayer Perceptron (MLP) Algorithm over a Health Disease dataset taken from the UCI machine learning repository. The MLP Back Propagation Algorithm was used to compute the depth of loss function in input data relative to all weights in the network. 13 attributes were mulled over for getting the objective.

The paper [9] discusses technology of Learning Vector Quantization (LVQA) Algorithm to provide a better, integrated, and accurate prediction of heart disease. The dataset used in this paper is a Heart disease dataset consisting of about 303 records and 13 attributes. The proposed architecture has four steps: Data sourcing; Performance Evaluation; Performance of Different Number of Neurons; Performance of Different Number of Training Epochs. The authors claim that this technology gives around 85.5% accuracy over the heart disease prediction.

2. METHODS OF DETECTION

2.1.1 Artificial neural network using health disease

M. Akhil Jabbar et al., (2021) designed classification of heart disease using artificial neural network and feature subset selection. An ANN is the model of the human brain. ANN is a supervised learning technique used for non linear classification. Clinical diagnosis is carried out by doctor's capability and patients were requested to take many number of diagnosis tests. However, all the tests fail to aim towards an effective diagnosis of disease. Feature subset selection is a preprocessing step that utilized to minimize the dimensionality and remove irrelevant data. PCA is employed for preprocessing and to minimize the number of attributes that minimizes the number of diagnosis tests.

Md. Osman Goni Nayeem et al., (2021) explained about Artificial Neural Network (ANN) that used for predicting heart disease. Feed-forward back propagation neural network algorithm with Multi-Layer Perceptron is implemented to identify infected or non-infected person. The Multi-Layer Perceptron with two hidden layer is applied to evaluate the medical diseases. This neural network system achieves high performance in predicting disease with minimal error.

2.1.2 Weighted fuzzy rules using diagnosis of health disease

P.K. Anooj (2021) presented a clinical decision support system with risk level prediction of heart disease using weighted fuzzy rules. Machine learning methods are designed to increase the knowledge from examples or raw data. Weighted fuzzy rulebased Clinical Decision Support System (CDSS) is designed for the diagnosis of heart disease finding knowledge from the patient's clinical data. The clinical decision support system is designed for the risk prediction of heart patients comprising of two phases. They are: computerized technique for the generation of weighted fuzzy rules and designing a fuzzy rule-based decision support system. The mining techniques are employed as attribute selection and attribute weight age method to attain the weighted fuzzy rules. Subsequently, the fuzzy system is planned appropriate to the weighted fuzzy rules and selected attributes.

2.1.3 Prediction and Diagnosis of Health Disease by using Data Mining Techniques

Boshra Bahrami and Mirsaeid Hosseini Shirvani (2020) designed prediction and diagnosis of heart disease by data mining technique. The significant application field is medical data mining. There are large amount of data available in healthcare however there is no efficient tool to determine hidden connections in data. Though, many people die of heart disease, data mining techniques in heart disease diagnosis are effective. Discovered Knowledge can help physicians in diagnosis of heart disease. The key aim is to calculate different classification techniques in heart disease diagnosis. Classifiers like J48 Decision Tree, K Nearest Neighbors (KNN), Naive Bayes (NB) and SMO are used to categorize dataset.

Divya Kundra and Er. Navpreet Kaur (2020) explained about data mining algorithms for the prediction of heart disease. The data mining technique is used to predict the information about heart disease that obtained the efficient outcome and reliable performance by using decision making. It will facilitate the medical practitioners to diagnose the disease in less time and analyze the possible difficulties. In order to estimate the most important risk factors of Heart Disease such as high blood cholesterol, diabetes, smoking, poor diet, fatness, hypertension and pressure. Data mining techniques employed to recognize the risk factors stage and helps to patients avoiding the heart failures.

2.1.4 Prediction of various health diseases

Tin Tin Su et al., (2020) designed the prediction of cardiovascular disease risk among low-Income urban dwellers in metropolitan kuala lumpur Malaysia. Framingham risk scoring (FRS) models are designed. An important determinant of the ten-year CVD risk is recognized by General Linear Model (GLM). The GLM models recognized the importance of education, occupation, and marital status in forecasting the CVD risk. Health care expenditure, illness related costs and loss of productivity because of CVD aggravated the existing situation of low-income urban population. The public health professionals and policy makers create substantial

effort to create the public health policy and community-based intervention to reduce the feasible high mortality and morbidity because of CVD between the low-income urban dwellers.

3. SURVEY PAPERS

Fast Health Disease Prediction Algorithm[FHDPA]: Authors of paper[12] proposed a new modular approach. The proposed Fast Health Disease Prediction algorithm (FHDPA) is designed and implemented to predict the severity level of heart diseases. The dataset used in this paper is a Cleveland Heart Disease dataset having 14 attributes, considered 10 attributes out of original dataset. They further modified the attributes information based on the expert views. The authors used WEKA tool to get decision rules for classification. FHDPA takes Cleveland dataset as input followed by a method that first scans the three major attributes which are age, resting blood sugar and cholesterol against the domain range, if three attributes fall under the normal range, then it falls under Risk level 0 and it exits. On the off chance that any of these values does not fall under normal range, at that point it additionally examines other attributes and yields the severity level of Heart Disease. The authors claim that this FHDPA calculation gave better outcome over different classifiers.

Prediction Based on Most Suitable Recommendation: The authors of this paper [13] proposed a model based on most suitable and appropriate values for early detection and correct diagnosis. The paper described their methodology, which was first divided into the age categories and assigned a fitness value to each attribute value. This value was an assumption based on medical knowledge; on the basis of these values they calculated fitness value for each record. Finally, they classified the tuple as per fitness value and calculated this value based on a defined formula. The proposed model is compared over Bayesian Classification and claims that the proposed model achieves better, accurate result. The authors considered only 5 main attributes for classification.

Online CHD Prediction Model: Authors of this paper [14] proposed an Online Adaptive Coronary Health Disease Risk Prediction Model (OCHD). Designed according to three risk factors like molecular structure, body system vital sign and Bio-energy symphony, this system calculates risk using static risk equation proposed in this paper. Malaysian dataset was used by the authors. The proposed model is tested among 120 subjects and validated with ECG and echo stress tests. The authors claim this model is compared with many other technologies and achieved highest score in their comparison. This technology is developed with in-depth study of medical factors that affect the heart.

C. Unsupervised Discrete Data mining Techniques in HDP

The papers dealing with the unsupervised learning algorithms are kept under this section.

Apriori: The authors of this paper [15] discussed the experiment that was executed using Association rules by using Apriori on medical data set to predict heart diseases. The author used this rule mining technology for rule extraction of different parameters on dataset; this rule mining based analysis was attempted by sorting information based on gender and different parameters which generated some set of rules. The authors used this technique to identify key factors behind this. The dataset used is Cleveland dataset having 14 attributes.

Association Classifier: Associative Classification is used to provide a better and accurate prediction of heart disease in [3][16]. Associative Classifiers (AC) discovers a rule subset with significant supports and higher confidence to build an automated classifier for prediction of previously unseen data classes. The paper does not give clarity on all components considered in the proposed architecture. The proposed model gives better and accurate results for heart disease prediction based on the Heart disease dataset having 7 attributes.

Stream Associative Classification Heart Disease Prediction (SACHDP): Two steps are applied to provide a better, integrated and accurate prediction of heart disease [17]. First, data stream having 14 attributes is stored in Prefix Streaming Tree structure, which is given as input in the form of landmark window in associative classification mining technique that generates rules. Second steps applied on the generated rules to Prune and arrange them to form a classifier. The paper does not give clarity on all components considered in the proposed architecture. The authors claim that combined technology gave better and accurate result over the heart disease prediction.

D. Comparison of Disrete Data Mining Techniques

This section compares the discrete data mining techniques using the various parameters such as number of attributes used, corresponding dataset, techniques and accuracy, which is given in Table III.

TABLE III: COMPARISON OF DISCRETE DATA MINING TECHNIQUES

Paper Citation	Attributes	Dataset	Techniques Used	Accuracy	Comments
[5]	NA	Cleveland	Naive Bayes	NA	No clarity on attributes and accuracy
[6]	8	NA	Naïve Bayes	NA	No clarity on dataset and accuracy
[7]	17	Korean	Decision Trees based Rule Induction	Male: 82.2%; Female: 83.5%	-
[8]	13	Cleveland	MLP	96.30%	-
[9]	13	Cleveland	Learning Vector Quantization Algorithm	85.55% for 18 neurons	No clarity on considered attributes
[10]	13	NA	Genetic Algorithm	Approx. 98%	No clarity on dataset
[11]	9	Real-life dataset	Series Recommendation Algorithm	75% - 100%	No clarity on dataset
[12]	14	Cleveland (modified)	Fast Heart Disease Prediction Algorithm	75.91%	No clarity on what other techniques they compared
[13]	5	NA	Own Method	Better than Bayesian Classifier	No clarity on dataset
[14]	NA	NA	CHD Prediction Model	96.20%	No clarity on dataset and attributes
[15]	14	Cleveland	Apriori Algorithm	NA	No clarity on accuracy
[16]	7	Various datasets	Association Classifier	NA	No clarity on accuracy
[17]	14	Dataset from UCI repository	Stream Associative Classification Heart Disease Prediction	94.94% for SACHDP	-

ANN and Particle Swarm Optimization (PSO): The authors of paper[22] have used combined technique of Artificial Neural Networks and particle swarm optimization to predict the Heart Diseases. Proposed methodology consist branching program, token generation and query. The paper does not give clarity on dataset or attributes considered for prediction. The authors did not provide any of accuracy measures of proposed technique for heart disease prediction.

Feed-forward neural network Particle Swarm optimization (PSO): The authors of this paper[23] discusses and presents the experiment that was executed with multilayer feed-forward neural network (MLFFNN), further optimized with particle swarm optimization(PSO) to predict the heart disease in early stage using the patient medical data. The main goal of PSONN was to achieve optimal training parameters such as hidden layer neurons, learning rate, momentum rate, transfer function in hidden layer and learning algorithm. The authors performed PSONN computationally efficient to predict heart disease. The performance analysis and compared with FFNN and SVM authors claim, result is accurate in PSONN compared to models to find best method giving accurate result. Developed SVM and FFNN technologies.

4. CONCLUSION

Increase in patients each year is a constant intimidation to each individual and a recurring problem for health authorities. Forecasting health disease based on various attributes of a patient can help authorities take measures to handle unexpected situation. Various proposals exist claiming evolution in accuracy for health disease prediction using data mining. This paper has compiled, tabulated, and analyzed these proposals with clear and appropriate taxonomy. The comparison of each category is based on various parameters such as number of attributes, usage of standard dataset, techniques used, accuracy achieved and specific remark. The taxonomy classifies them based on usage of number of data mining technique into discrete (one) or integrated (multiple) approach at first level. Discrete approach performs well but integrated approach gives promising results in health disease prediction compare to discrete approach.

REFERENCES

[1] Kaur, Beant, and Williamjeet Singh. "Review on heart disease prediction system using data mining techniques." International journal on recent and innovation trends in computing and communication, Vol.2, No.10 , 2014, pp. 3003-3008.

[2] K.Sudhakar and Dr. M. Manimekalai.," Study of Heart Prediction using Data Mining". IJARCSS, Vol.4,No. 1, January 2014, pp.1157-1160

[3] Karthikeyan, T., B. Ragavan, and V. A. Kanimozhi. "A Study on Data mining Classification Algorithms in Heart Disease Prediction." Int. J. Adv. Res. Comput. Eng. Technol, Vol.5, No.4, 2016, pp.1076-1081.

[4] Banu, NK Salma, and Suma Swamy. "Prediction of heart disease at early stage using data mining and big data analytics: A survey." Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference, IEEE, 2016, pp. 256-261.

[5] Baiju, B. V, and RJ Remy Janet, "A Survey on Heart Disease Diagnosis and Prediction using Naive Bayes in Data Mining", INPRESSCO- International Journal of Current Engineering and Technology, Vol.5, No.2 ,April 2015, pp. 1034-1038

[6] Rana, Ruchika, and Jyoti Pruthi. "Heart Disease Prediction using Naive Bayes Classification in Data Mining." IJSRDInternational Journal for Scientific Research and Development, Vol.2, No.05, 2014 pp.2321-0613.

[7] Kim, Jae-Kwon, Eun-Ji Son, Young-Ho Lee, and Dong-Kyun Park. "Decision tree driven rule induction for heart disease prediction model: Korean National Health and Nutrition Examinations Survey V-1.", IT Convergence and Security 2012, Springer, Dordrecht, Vol.2, No.5, 2013, pp. 1015-1020.

[8] Durairaj, M., and V. Revathi, "Prediction of Heart Disease Using Back Propagation MLP Algorithm." International Journal Of Scientific and Technology Research, Vol.4, No.08, 2015, pp.236-239.

[9] Sonawane, Jayshril S., and D. R. Patil. "Prediction of heart disease using learning vector quantization algorithm." IT in Business, Industry and Government (CSIBIG), 2014 Conference on. IEEE, 2014, pp.1-5.

[10] Dhokley, Waheeda, Tahreem Ansari, Naeema Fazlani, and Heena Mohd Hafeez. "New Improved Genetic Algorithm for Coronary Heart Disease Prediction." International Journal of Computer Applications Vol.136, No. 5, 2016, pp.34-37.

[11] Lafta, Raid, Ji Zhang, Xiaohui Tao, Yan Li, and Vincent S. Tseng. "An intelligent recommender system based on shortterm risk prediction for heart disease patients." Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference on, Vol. 3, 2015 , pp. 102-105.

[12] Saxena, K., and Sharma, R. "Modular Approach for Heart Disease Prediction". Indian Journal of Science and Technology, Vol. 9, No. 45, 2016, pp.1-5.

[13] Saiyed, Kasim Ali, and Vijay Kumar Verma. "Prediction for Heart Disease Problem Based on Most Suitable Recommendation." International Journal ,Vol.1, No. 7, 2016, pp.6-10.

[14] Lam, J., Supriyanto, E., Yahya, F., Satria, M.H., Kadiman, S., Azan, A. and Soesanto, A.,"Online Adaptive Coronary Heart Disease Risk Prediction Model." MATEC Web of Conferences ,Vol. 125, 2017, pp. 02071.

[15] Said, Ibrahim Umar, Jamila M. Muhammad, and Manoj Kumar Gupta. "Intelligent Heart Disease Prediction System by Applying Apriori Algorithm." International Journal Vol.5, No.9 ,2015, pp.887-891

[16] Jabbar, M. Akhil, Bulusu Lakshmana Deekshatulu, and Priti Chandra. "Knowledge discovery using associative classification for heart disease prediction." Intelligent informatics, Springer, Berlin, Heidelberg, 2013. pp. 29-39

[17] Lakshmi, K. Prasanna, and C. R. K. Reddy. "Fast rule-based heart disease prediction using associative classification mining." Computer, Communication and Control (IC4), International Conference, 2015, pp. 1-5. IEEE, 2015.

[18] Akila, S., and S. Chandramathi. "A Hybrid Method for Coronary Heart Disease Risk Prediction using Decision Tree and Multi Layer Perceptron." Indian Journal of Science and Technology Vol.8, No.34 ,2015, pp.1-5.

[19] Shinde, Aditya A., Rahul M. Samant, Atharva S. Naik, Shubham A. Ghorpade, and Sharad N. Kale. "Heart Disease Prediction System using Multilayered Feed Forward Neural

[20] Dewan, Ankita, and Meghna Sharma. "Prediction of heart disease using a hybrid technique in data mining classification." Computing for Sustainable Global Development (INDIACom), 2nd International Conference ,IEEE,2015, pp. 704-706.