A Survey on Diagnosis of Heart Diseases using Data Mining Techniques

Tashrifa Shahid
Department of Computer Science & Engineering
Prime University

Ferdousi Barira
Department of Computer Science & Engineering
Prime University

ABSTRACT
Data mining tools are effectively used in disease diagnosis which helps health professional. From health sector a large number of data are collected, classification tools are applied on these data and discover new pattern. In this paper, heart diseases have been chosen for diagnosis and classification. An extensive analysis is performed on some popular data mining methods by using a large number of datasets in this work. To understand the major data mining techniques and select the suitable category of algorithms, the analysis result will help for heart disease analysis. Decision tree has successfully used in different research to predict disease. In this research, decision tree is applied to classify hypertension disease.

General Terms
Pattern Recognition, Algorithms

Keywords
Decision Tree, Weka tool, Naive Bayes, Cardiovascular disease,SVM

1. INTRODUCTION
The word heart disease applies to a wide range of illnesses that affect the circulatory system, which consists of heart and blood vessels. It is intended to deal only with the condition normally known as "Heart Attack" and the factors, which lead to such condition. Cardiomyopathy and Cardiovascular disease are some types of heart diseases.It is observed that CHD is the cause considered for 17.7 million deaths every year and more than twenty-four million ratio of people anticipated passing from cardiovascular sickness by the year of 2030 (Kinge & Gaikwad, 2018) CHD dominates other diseases with its severe effects on a person's wellbeing worldwide (Wilson et al., 1998)[28]. Anabrupt blockage of a coronary artery, usually due to a blood clot results in a heart attack. The purposes of this paper is to evaluate the distinct predictive/ descriptive data mining approaches proposed in current years for the diagnosis of heart disease.

2. BACKGROUND
C4.5 Decision Tree was used by Andreeva in the diagnosis of heart disease that produced accuracy of 75.73% (Andreeva 2006). The optimized K-Means algorithm was used by Shimpli et al. 2020 with precision 91% and recall of the model was 75.83%. In 2017 NN-based CHD risk prediction used by Kim and Kangwhere feature correlation analysis evaluated highest accuracy (82.51%) in a CHD prediction [29]. The best-performing data mining technique that the heart disease prediction model developed using the identified significant features achieves an accuracy of 87.4% was analyzed by Amin et al. in heart disease prediction (Amin, Chiam et al., 2019). SVM-Radial bias kernel technique was improved by Karthi keyan et al. which was produced accuracy of 89.9% in heart disease (Karthi keyan et al., 2020). K-Means and Artificial Neural Network techniques were combined by Amita Malav, Kalyani Kadam and Pooja Kamat to achieve higher prediction accuracy[22].

3. METHODOLOGY
Because of the limitations of resources and the characteristics of the paper itself, the main methodology used for this paper was throughout the survey of journals and publications in the areas of medicine, computer science and engineering. A range of current publications are followed in this research.

3.1 Heart disease prediction using data mining
Three different supervised machine learning algorithms: Naive Bayes, K-NN, Decision List, SVM algorithm have been used for analyzing the dataset in [4]. Tanagra tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. These mentioned algorithms were applied to predict the accuracy of heart diseases.

3.1.1 Naïve bayesclassifier
Naïve Bayes is a classification algorithm which follows the probability theory to detect most considerableprobable classifications. It categorize by applying Bayes theorem with strong (naive) independence predictions. In simple word, a this classifier assumes that the existence (or unavailability) of a certaincharacteristics of a class is independent to the existence (or unavailability) of any other characteristic. According to the pinpointqualities of the probability model, in a supervised learning settingthis classifiers can be trained more effectively.

3.1.2 K-Nearest neighbor algorithm
The k-nearest neighbor‘s algorithm (k-NN) is a method for classifying objects based on closest training data in the feature space. The similartechnique can be used for regression, by allocating the property value for the item to be the mean of the values of its k nearest neighbors. It can be helpful to weight the assistances of the neighbors, so that the closer neighbors submit more to the mean than the more far ones.

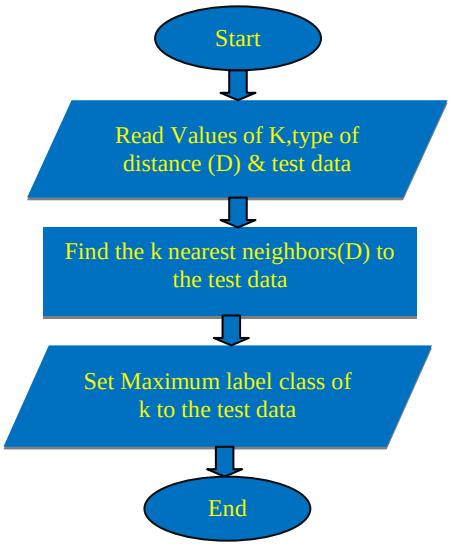

Fig 1: The flowchart of K- Nearest Neighbor Classifier Procedure [26]

3.1.3 Support vector machine

According to the hypothesis of statistical learning and the standard of structural reduction of risk, support vector machine can perform pattern recognition and regression. Support vector machine applied in weka data mining tools is Sequential Minimal Optimization that is an algorithm for effectively solving the optimization problem that appearsthrough the raining of Support Vector Machines.

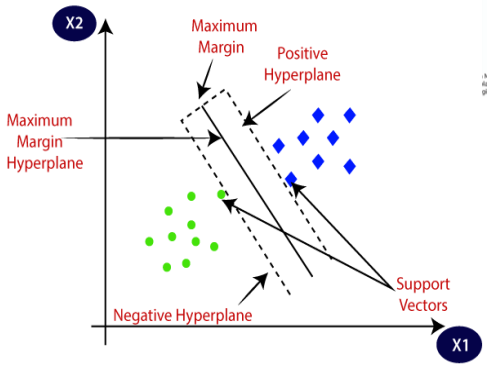

Fig 2: SVM Classified Two Different Categories Using Hyperplane [27]

3.1.4 Decision tree

Decision tree is more alike to the flowchart in which every non-leaf nodes denotes a test on a specific attribute and each branch contributes an outcome of that test and every leaf node have a class label. The top most labeled node in the tree is called root node. By applying Decision Tree, decision makers can find best alternative and traversal from root to leaf indicates exceptional class division based on highest information gain.

4. PERFORMANCE STUDY OF ALGORITHMS

Accuracy for different classification method with different attributes values.

4.1 Description of Dataset

In reference [1] they have used the dataset having the following attributes

- id: identification number of patient
- age: age of patient in year,
- sex: sex of patient (1=male; 0 = female),
- painloc: patient's chest pain location (1 = substernal; 0 = otherwise),
- pain_exer (1 = provoked by exertion; 0 = otherwise),
- rel_rest (1 = relieved after rest; 0 = otherwise),
- cp: chest pain type
  - Value_1: typical angina
  - Value_2: atypical angina
  - Value_3: non-anginal pain
  - Value_4: asymptomatic
- trestbps: patient's resting blood pressure
- chol: patient's serum cholestoral
- famhist: patient's family history of coronary artery disease (1 = yes; 0 = no)
- rest_ecg: patient's resting electrocardiographic results
  - Value_0: is normal
  - Value_1: having ST-T wave is abnormal (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value_2: displaying probable or definite left ventricular hypertrophy by Estes' criteria ekgmo (month of exercise ECG reading)
- ekgmo (month of exercise ECG reading)
- thal_dur: period of exercise test in minutes
- thal_ach: highest heart rate achieved
- Thal_rest: resting heart rate
- num: diagnosis of heart disease (angiographic disease status)
  - Value_0: < 50% narrow diameter
  - Value_1: > 50% diameternarrowing
- (in any major vessel: attributes 59 through 68 are vessels)

In reference [3] they have used the dataset having the following attributes. Data source hsa total 909 records with 15 healthfactors which were obtained from the Cleveland Heart Disease database.

- Diagnosis(value_0: <50% diameter narrowing(no heart disease); value_1: > 50% diameter narrowing(heart disease))
- Key attribute:
  Patient_ID- Patient's identification number
- Input attribute:
  - Sex(value_1: Male; value_0: Female)
  - Chest pain type (value_1:typical type I angina, value_2: typical type angina, value_3:non angina pain; value_4:asymptomatic)
  - Fasting Blood Sugar(value_1: >120 mg/dl; value_0: <120 mg/dl)
  - Resteeg- resting electrographic result(value_0: normal; value_1: having ST-T wave abnormality; value_2: showing probable or definite left ventricular hypertrophy)
  - Exang- exercise induced angina(value_1:yes; value_0: no)
  - Slope- the slope of the peak exercise ST segment(value_1:unsloping; value_2: flat; value_3: downsloping)
  - CA- number of major vessels colored by floursopy(value 0-3)

- Thal (value_3: normal; value_6: fixed defect; value_7: reversible defect)
- Test Blood Pressure of patient(mm Hg on admission to the hospital)
- Serum Cholesterol(mg/dl)
- Thal_ach – highest heart rate achieved
- Old_peak- ST depression induced by exercise relative to rest
- Age of patient in Year

Table 2 contains the features of reference [8]. Data do not need of integration operations because it is collected from one resource.

Table 1: List of attributes [8]

Sl No	Nameof patient	Possible Values
1	Age of Patient	NUMERIC VALUES
2	Chest_pain_ typeS	ASYMPT, ATYP_ANGINA, NON_ANGINAL,TYP_ ANGIN
3	rest_bp_ress	NUMERIC VALUES
4	blood_sugar	TRUE, FALSE
5	rest_electro card	Normal,left_vent_hyper st_t_wave_abnormality
6	hightest_he art_rate	NUMERIC VALUES
7	exercice_an gina	YES, NO
8	Disease	NEGATIVE, POSITIVE

4.2 Description of Tools

In reference [2] they have used Tanagra dataset. The Tanagra project allows to evaluateeither real or synthetic data.
In this study, to apply the data mining algorithms WEKA was used. In this part experimental results from implementation of different classification algorithms, j48 decision tree, Naive Bayes, KNN and SMO on heart disease datasets are evaluated and compared.

4.3 Performance Analysis and Evaluation

Secondary values of different classifications are given in table 2. Considering these values the precision of algorithms are calculated and analyzed. Classifier has 14 attributes. According to evaluation time of calculation and the error rates, performance can be determined.

Table 2. Performance Study of Algorithm [1]

Algorithm Used	Accuracy	Time Taken
Naive Bayes	52.33%	609ms
Decision List	52%	719ms
KNN	45.67%	1000ms

Experimenting with a training dataset, Naive bayes algorithm results in less error ratios than decision list and k-NN algorithm.

The records were divided equally into two datasets: training dataset with 455 records and testing dataset with 454 records. The summary of the results of all three models are shown in Table 3. SinceNaïve Bayes has the highest percentage of correct predictions (86.53%) for patients with heart disease, it appears to be most effective, followed by Neural Network (with a difference of less than 1%) and Decision Trees. For predicting patients with no heart disease decision trees appears to be most effective (89%) compared to the other two models [3].

Table 3. Accuracy of Different Classifiers [3]

Techniques	Accuracy
Naive Bayes	86.53 %
Decision Tree	89%
ANN	85.53 %

Table 4 shows accuracy for different classification method with 13 input attributes & 15 input attributes values [11].

Table 4: Comparison of data mining techniques

Classification Techniques	Accuracy with	
	13 attributes	15 attributes
Naive Bayes	94.44	90.74
Decision Trees	96.66	99.62
Neural Networks	99.25	100

Table 5. Comparison between accuracies of different techniques [5]

Techniques	Accuracy
Naïve Bayes Decision tree Neural network Kernel density [16]	78.563% 75.738% 82.773% 84.449%
Naïve bayes Decision tree Neural network [17]	95% 94.93% 93.54%
Naïve bayes Decision tree [18]	62.03% 60.40%
Naïve bayes KNN	52.33% 45.67%
Decision list [19]	52%
Naïve bayes One dependency augmented Naïve Bayes classifier [20]	84.14% 80.46%
Genetic with decision tree Genetic with Naïve Bayes Genetic with classification via clustering [21]	99.2% 96.5% 88.3%

From the Table 5, observe that Naive Bayes's accuracy is highest as contrast to all other classification techniques. For this reason, this technique is used in proposed system for prediction of heart disease using WSN.

5. CONCLUSION

Since worldwide one of the leading causes of death isheart disease, the early prediction of heart disease is essential. This paper reviewed some Heart Disease classification system. In this work different techniques and data mining classifiers are defined which has merged in current years for diagnosis of heart disease efficiently and effectively. The surveydepict

thatall the papers use different technologies with different number of attributes. So, accuracy varies withusing different technologies to classify the attributes.

6. REFERENCES

[1] Asha Rajkumar, Mrs. G.SophiaReena, "Diagnosis Of Heart Disease Using Data mining Algorithm", Global Journal of Computer Science and Technology, P a g e |38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[2] Mai Shouman, Tim Turner, Rob Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients", Proceeding of the 9th Australian Data Mining Conference (AusDM' 11), Ballarat, Australia.

[3] JyotiSoni, Ujma Ansari, Dipesh Sharma, SunitaSoni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[4] Asha Rajkumar, G.SophiaReena, Diagnosis Of Heart Disease Using Data mining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[5] PrachiJambhulkar, VaidehiBaporikar, "Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network", International journal of Computer Science and Application, Vol. 8, No.1, Jan-Mar 2015.

[6] Apte&S.M. Weiss, Data Mining with Decision Tree and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsap tewe issue_with_cover.pdf,(1997).

[7] K. Thenmozhi, P.Deepika, "Heart Disease Prediction Using Classification with Different Decision Tree Techniques" International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014.

[8] BoshraBahrami, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February – 2015.

[9] Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D.,Jayaraman, V. K., and Balaji, P. V."A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on over expression in escherichiacoli,"Bioinformatics, 2006.

[10] Platt, John, "Sequential minimal optimization: a fast algorithm for training support vector machines,"Technical Report Microsoft Research, 1998.

[11] Chaitrali S. Dangare ,Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[12] SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

[13] Mrs.G.Subbalakshmi "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering (IJCSE) ISSN: 0976-5166, Vol. 2 No. 2 Apr-May 2011, pp.170-176.

[14] PrachiJambhulkar, VaidehiBaporikar, "Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network", International Journal of Computer Science And Applications,Vol. 8, No.1, Jan-Mar 2015, ISSN: 0974-1011.

[15] VikasChaurasia and Saurabh Pal "Early Prediction of Heart Diseases Using Data Mining Techniques", Caribbean Journal of Science and Technology, 2013, ISSN 0799-3757, Vol.1, pp.208-217.

[16] Andreeva, P. "Data Modeling and Specific Rule Generation via Data Mining Techniques". International Conference on Computer Systems and Technologies - CompSysTech, 2006.

[17] SellappanPalaniappan, RafiahAwang "Intelligent Heart Disease Prediction System Using Data Mining Techniques",978-1-4244-1968- 5/08/$25.00 ©2008 IEEE.

[18] Sitar-Taut, V.A. "Using machine learning algorithms in cardiovascular disease risk evaluation", Journal of Applied Computer Science & Mathematics, 2009.

[19] Rajkumar, A. and G.S. Reena "Diagnosis of Heart Disease Using Data mining Algorithm", Global Journal of Computer Science and Technology, 2010. Vol. 10 (Issue 10).

[20] Srinivas, K., "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining Techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[21] M. Anbarasi "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10),pp.5370-5376, 2010.

[22] KalyaniKadam, AmitaMalav," A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K-means", International Journal of Engineering and Technology 2017.

[23] S. Shilna and E. Navya, " Heart disease forecasting system using k-means clustering algorithm with PSO and other data mining method," International Journal On Engineering Technology and Sciences (IJETS), ISSN(P): 2349-3968, ISSN(O): 2349-3976, Vol 3, Issue 4, April 2016.

[24] K.R. Lakshmi, M. V. Krishnaand P Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability", International Journal of Scientific and Research Publications, ISSN 2250-3153, Vol.3, Issue.6, June 2013.Survivability,"International Journal of Scientific and Research .

[25] K. Solanki, P. Berwal and S. Dalal, "Analysis of application of data mining techniques in healthcare," International Journal of Computer Applications, August 2016.

[26] Mohammad Bazmara, SaniaVahedianMovahed, Samira Ramadhani, "KNN Algorithm for Consulting Behavioral Disorders in Children", Journal of Basic and Applied Scientific Research, 2013.

[27] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[28] Nancy Masih, Sachin Ahuja, "Prediction of Heart Diseases Using Data Mining Techniques: Application on Framingham Heart Study," International Journal of Big Data and Analytics in Healthcare Volume 3 , Issue 2, July-December 2018.

[29] Jae Kwon Kim, Sanggil Kang, "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis," Journal of Healthcare Engineering, Volume 2017.

[30] Mohammad Shafenoor Amin, Yin Kia Chiam, KasturiDewiVarathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, Volume 36, March 2019.

[31] KarthikeyanHarimoorthy, MenakadeviThangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system," Journal of Ambient Intelligence and Humanized Computing, 2020.