IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Appropriate Attributes And Stacking

Shubhangi S. Jadhav
S.G.G.S.IE&T,Nanded(MH)

Dr.M.V.Vaidya
Associate Proffesor atS.G.G.S.I.E.&T.Nanded(MH)

Abstract

Cardiovascular disease or heart disease is the world's largest cause of weariness and death. Predicting cardiac disease before hand is most important in the analysis clinical data. A large amount of data is present in health care industry ,that data can be converted into information , with the help of that information predictions can be made. Several researchers have applied data mining technique to predict heart disease. It is very important to select appropriate set of attributes and data mining algorithm to enhance predictive accuracy. This research is done to select appropriate attributes and data mining algorithms that can enhance predictive accuracy of heart disease. Different sets of attributes and classification techniques were used to develop prediction model. K-NN, Logistic Regression, Support Vector Machine,Decision Tree, Naïve Bayes, Vote(Hybrid technique with naïve bayes and logistic regression) and Stacking(A hybrid technique with naïve bayes, logistic regression as base learners and support vector machine as model learner).This study shows that the cardiac disease prediction model was developed by using appropriate attributes and best results of data mining technique(stacking) got an accuracy of 85.71% in predicting cardiac disease.

Keywords:Data mining,Attribute selection,Classification technique,Prediction model

Introduction

"Cardiovascular diseases are the number one cause of death world wide. According to 2016 World Health Organisation(WHO) survey 17.9 millions people died to heart disease. CVDs are the no one cause of death globally; more people die annually from CVDs than from any other cause"[9]. If we can predict heart disease and caution ahead then we can prevent many of deaths. "Many data mining techniques are used for identifying and extracting useful information from the clinical data set with minimum user inputs and efforts"[11].In the last few years researchers analysed many ways of applying data mining in health care system in order to obtain an effective prediction of CVDs. The performance of data mining algorithm many differs on algorithm used and attributes selected. "The medical data sets in the health care industry are redundant and inconsistent. It is quite hard to use data mining technique without pre processing"[10] . The data mining performance varies according to techniques used and attributes selected. Data mining technique is hard to use without appropriate and adequate preparations.

"Data redundancy and inconsistency in raw data set affects the predicted outcome of algorithms.As a result to apply the machine learning algorithms to its full potential, an effective preparation is needed to pre process the data set"[1]. "The unwanted features can reduce the performance of data mining techniques"[2]. "Therefore along with data preparation, a proper feature selection method needed to achieve high Accuracy in heart disease prediction using significant features and data mining techniques"[10].

"Despite feature selection is an important as selecting suitable technique, researchers still agonizing in combining appropriate data mining technique with a proper set of features. There is an expectation to diagnose the CVDs with high accuracy, but it is not easy to achieve"[3]. Also a set of appropriate attributes will definitely improve the accuracy of prediction. This shows identifying appropriate attributes is important to acquire that aim.

"The performance of data mining techniques used in predicting cardiovascular diseases is greatly reduced without a good combination of key features and also the improper use of the machine learning algorithms ,thus it is important to identify the correct set of appropriate attributes that works well with the best performing algorithm"[5]. This research is mainly on finding correct data mining technique with appropriate attributes that will perform well in predicting heart disease. Though it is hard to choose the correct technique and select appropriate attributes . Current studies has shown that data mining techniques are used in predicting heart disease are insufficient , "a proper examination require to identify the significant features and data mining techniques that will improve the performance. A proper evaluation and comparison to test the different combination of features together with data mining techniques are yet to be focussed "[8]. Thus identification of appropriate attributes and data mining technique is needed to produce accurate results.

This research was done to identify important attributes and data mining algorithm to predict heart disease. The heart disease dataset were collected from UCI machine learning repository. Eight classification techniques were applied to create prediction models. By using result of experiment, eight important attributes and four best data mining techniques were selected.

Literature Survey

Heart disease is the common health problem these days. Many of studies have already done on heart disease prediction. Researchers are trying to improve the accuracy and performance of model. According to (Y K Singh et al ,2017), data processing is more important as insufficient and incorrect data may reduce the accuracy of data mining technique.

The effectiveness of data mining depends on the technique used and attribute select most of the dataset provided by health care industry are inconsistent and redundant data inconsistency and redundancy in the data set can affect the result of algorithms, according to Kavita and Kannan.

Moreover, "inappropriate attributes may lower the performance of the data mining technique too"[2] therefore along with data processing, an appropriate attribute selection technique is needed to get great accuracy in heart disease prediction using appropriate feature and data mining technique.

Selection of attributes is equally important as the correct technique, Researchers are still troubling in combining appropriate set of attribute and data mining technique. There is an expectation that the heart disease will be detected with best possible accuracy but it is hard to achieve[3] . In fact a combination of appropriate attributes certainly would increase prediction accuracy. It indicates that how much it is necessary to specify appropriate attributes to achieve that goal.

"The proper use of machine learning algorithms and a combination of appropriate attributes can improve the performance of data mining techniques used in predicting heart disease "[5].Thus it is most important to select appropriate set of attributes that will work great with best performing algorithms. Previous studies have

also shown that data mining algorithms used in prediction of heart disease are inadequate and a careful analysis is needed to choose the important attributes and data mining techniques may improve the accuracy.

"A thorough assessment and comparison to test the various subsets of attributes together with data mining technique are yet to be focussed "[8]. According to M S Amin et al, 2018 proper selection of data mining techniques and appropriate attributes to make the prediction of heart disease accurate and acceptable is needed[10].

2 Dataset

"The heart disease data were collected from UCI machine learning repository.Cleveland dataset was used for this research because it contains complete records. The dataset has total 303 records and 76 attributes. But the dataset provided in repository has 14 attributes"[6]. There are total 13 attributes for feature and 1 attribute for result of prediction.

Table 1 Dataset description

Attribute	Type
Age	Numeric
Sex	Nominal
Cp	Nominal
Trestbps	Numeric
Chol	Numeric
Fbs	Numeric
Restecg	Nominal
Thalach	Numeric
Exang	Nominal
Oldpeak	Numeric
Slope	Nominal
Ca	Numeric
Thal	Nominal
Num	Nominal

3.Methodology

In this study the experiment was conducted by using RapodMiner Studio. As visual design environment is easy to use to build predictive model. The dataset of heart disease was importded into RapidMiner Studio. After that pre processing was performed in this step process of data mining starts. Third step is attribute selection. In fourth step classification models were applied on selected attributes , the log of performance of respective algorithms was maintained. Output of result is shown after completing entire process.

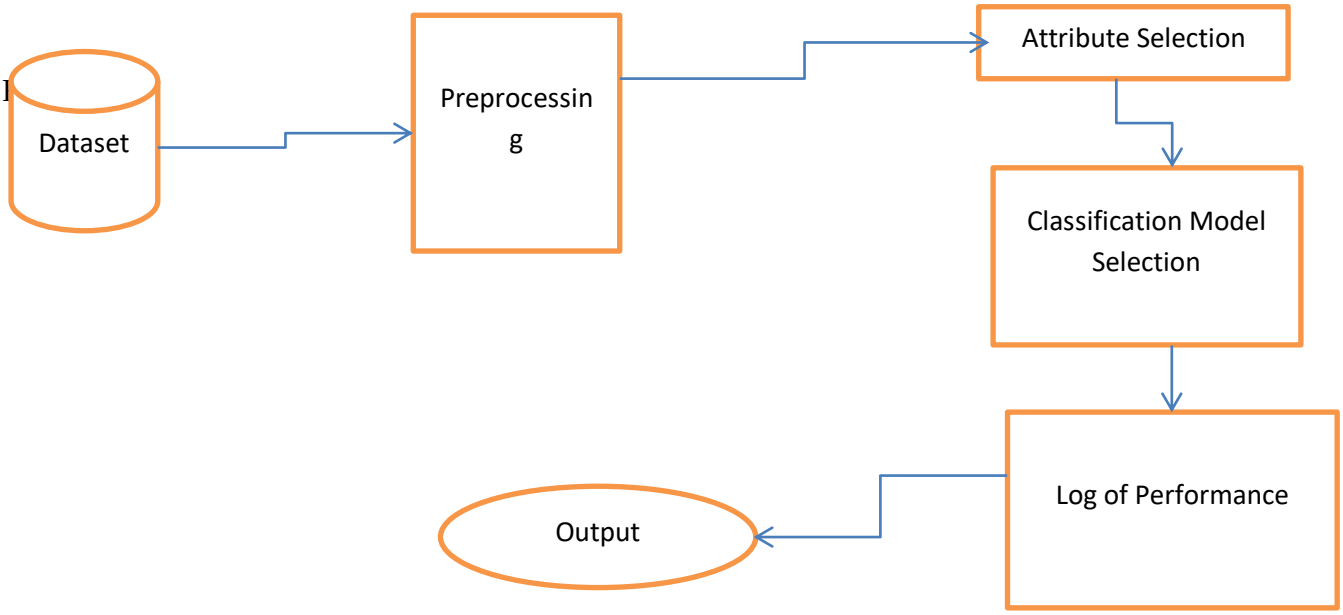

Fig. Workflow of experiment

3.1 Pre-processing

The data were pre-processed after collection. The values of result attribute in the dataset was converted from multiclass values to binary values. Cp , Thal, slope and Num were converted from multiclass values to binary values

3.2 Attribute selection

There are total 14 attributes present in dataset from which 13 are used in prediction of heart disease and 1 for result. In this study set sof attributes were selected to be used with 8 classification techniques, to create the classification model. The techniques were K-Nearest Neighbour, Decision tree, Naïve Bayes, Logistic Regression, Vote, Support Vector Machine, Neural Network and Stacking(a hybrid algorithm with logistic regression , SVM as base learners and stacking as model learner) .To select appropriate attributes ,all possible combination attributes with all selected techniques were tested.

$$2^n - (n^2+n/2+1)$$

Where n is total number of attributes used to generate the subses of combination, which is 13.

3.3 Selection of classification Technique

The model were created with 8 appropriate classification techniques: Logistic Regression, Waive Bayes, Decision Tree, Support Vector Machine, Neural Network, K-NN, Vote and Stacking(i. e. a hybrid technique with logistic regression and SVM as base learner)K-folds cross validation technique) was used to validate the performance of the modes. In this study data set was divided into 10 subsets were used as testing sets and remaining one subset as training set.

3.4 Performance Technique

The performance of developed model were measured using accuracy, precision and recall. Accuracy is percentage of correctly predicted instances among all precision is percentage of correctly predicted positive observations. Recall is percentage of actual positives a model correctly identified.

4.Results Analysis

Table 2 The accuracy ,precision and recall obtained by each data mining algorithm with 13 attributes.

Technique	Accuracy	Attribute	Precision	Recall
Decision Tree	69.64%	Thal,trestbps,chol,fbs,restecg,cp,thalach,exang,oldpeak,slope,Ca, age, sex	68.22%	63.31 %
KNN	56.11%	Thal,trestbps,cp,chol,fbs,restecg,thalach,exang,oldpeak,slope,Ca,sex ,age	58.96%	48.92 %
Logistic Regression	79.21%	Thal,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope, cp, sex,Ca, age,sex	79.88	75.54 %
Naïve Bayes	82.18%	Thal,trestbps,chol,fbs,restecg,cp,thalach,exang,oldpeak, sex, slope,Ca, age	79.86%	83.13 %
Neural Net	80.53%	Thal,trestbps,chol,fbs,restecg,thalach,cp,exang,oldpeak,slope,Ca, age, sex	79.66%	74.10 %
SVM	83.50%	Thal,trestbps,chol,fbs,restecg,sex,thalach,exang, oldpeak,slope, cp,Ca, age	80.32%	73.38 %
Vote	84.16%	Thal,trestbps,chol,sex,fbs,restecg,thalach,exang,cp,oldpeak,slope,Ca, age	81.87%	76.26 %
Stacking	82.42%	Thal,trestbps,chol,fbs,cp,restecg,thalach,exang,oldpeak,slope,Ca, age, sex	83.67%	80.95 %

Table 3 The accuracy ,precision and recall obtained by each data mining algorithm with 9 attributes.

Technique	Accuracy	Attribute	Precision	Recall
Decision Tree	73.27%	Fbs,cp,restecg,sex,oldpeak,slope,ca,thal,exang	71.97%	68.35%
KNN	80.86%	Fbs,cp,restecg,sex,oldpeak,slope,ca,thal,exang	82.40%	74.10%
Logistic Regression	83.50%	Fbs,cp,restecg,sex,oldpeak,slope,ca,thal,exang	79.14%	83.697%
Naïve Bayes	82.18%	Fbs,restecg,cp,sex,oldpeak,slope,ca,thal,exang	82.44%	77.70%
Neural Sets	82.84%	Fbs,restecg,sex,cp,oldpeak,slope,ca,thal,exang	83.72%	77.70%
SVM	83.83%	Fbs,restecg,sex,oldpeak,cp,slope,ca,thal,exang	89.47%	73.38%
Vote	83.17%	Fbs,restecg,sex,oldpeak,slope,cp,ca,thal,exang	86.07%	75.54%
Stacking	84.62%	Fbs,cp,restecg,sex,oldpeak,slope,ca,thal,exang	85.00%	80.95%

Table 3 The accuracy ,precision and recall obtained by each data mining algorithm with 8 attributes.

Technique	Accuracy	Attribute	Precision	Recall
Decision Tree	75.91%	oldpeak, fbs,exang,sex,slope,ca,thal,cp	76.16%	69.06%
KNN	80.20%	cp,oldpeak,fbs,exang,sex,slope,ca,thal	79.70%	76.26%
Logistic Regression	84.49%	oldpeak,fbs, cp,exang,sex,slope,ca,thal	85.94%	79.14%
Naïve Bayes	82.18%	oldpeak, fbs,exang, cp,sex,slope,ca,thal	82.44%	77.70%
Neural Sets	83.50%	Oldpeak,fbs, cp,exang,sex,slope,ca,thal	84.50%	78.42%
SVM	83.50%	oldpeak,fbs,exang,sex,slope, cp,ca,thal	88.70%	73.38%
Vote	83.83%	oldpeak,fbs,exang,sex,slope,ca,thal ,cp	88.14%	74.82%
Stacking	85.61%	oldpeak, fbs,exang,sex,slope,ca, cp,thal	87.18%	80.95%

5.Evaluation

The Accuracy achieved through prediction models built using 8 appropriate attributes and best performing three classification techniques are as follows

Table 4 Comparison between best techniques

	Stacking	Vote	SVM
Accuracy with 13 features	82.42%	84.16%	83.50%
Accuracy with 9 features	84.62%	83.17%	83.83%
Accuracy with 8 features	85.17%	83.83%	83.50%

Table summarizes the accuracy of models the analysis produced. Table shows 13 attributes, 9 attrbutes and 8 appropriate attributes. Accuracy of prediction model developed using 8 appropriate attributes is improved than the models built using all 13 attributes. The highest accuracy(85.71%) was achieved by stacking using 8 appropriate features.

6.Discussion

The result in tables shows that the selected 8 appropriate attributes have enhanced the accuracy of data mining techniques. 7 out of 8 selected features are clinical attributes and one attribute i. e. sex related to personal information of patient. This shows clinical features are more important than personal information in heart disease prediction using data mining technique.

As per the table stacking have achieved highest Accuracy of 85.71% with 8 appropriate attribute figures shows overviews of the model for prediction heart disease proposed.

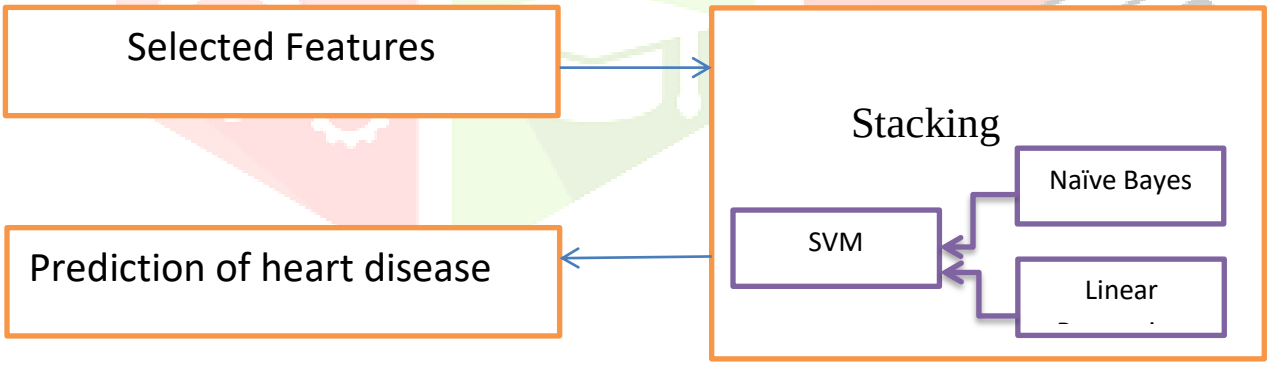

Fig. Workflow of proposed model

This method was based on model proposed to help to predict low cost cardiac disease in patients.

6.1 Comparison of proposed model

Table shows comparison between existing studies and proposed model. Proposed model achieved highest accuracy i.e 85.71% with appropriate set of features.

Table 5 Comparison between existing studies and proposed model

	Technique used	Accuracy
Proposed model	Stacking with logistic regression, naïve bayes and SVM	85.71%
M.S.Amin et al.(2018)	Vote	83.83%
Poul et al.(2016)	Neural Network with fuzzy	80%
Nahar et al.(2013)	Naive Bayes	69.11%
Shouman et al.(2013)	Decision Tree with Gain Ratio	84.10%

Conclusion

Heart disease is no one cause of death worldwide. It is very important to discover it as soon as feasible to prohibit deaths.

This study was done to improves the accuracy of prediction by combining appropriate attributes and the best classification modelling technique . The experiment was carried out using UCI cleveland data collection to determine appropriate attribute and techniques for data mining. The eight selected attributes are Sex, CP, Fbs, Rest ECG, oldpeak, Slope, Thal, Ca. Stacking, Vote and SVM are the best performing data mining techniques with maximum accuracy . The assessment results verifies that the selected attributes are appropriate.

Future Scope

This research can be done by using real time data set. Further research can be done by combining different data mining techniques. New attribute selection method can be applied to improve accuracy of prediction model.

References

[1] Kavitha, R., Kannan, E., 2016. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), pp. 1–5.

[2] Paul, A.K., Shill, P.C., Rabin, M.R.I., Akhand, M.A.H., 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. (ICIEV). In: 5th
International Conference on Informatics, Electronics and Vision. IEEE, pp.145-150

[3] Shouman, M., Turner, T., Stocker, R., 2013. Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng. 20 (1).

[5] Dey, A., Singh, J., Singh, N., 2016. Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal
component analysis. Analysis 140 (2), 27–31.

[6] Dua, D., Karra Taniskidou, E., 2017. UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA http://
archive.ics.uci.edu/ml.

[7]Singh,Y.K.,Nikhil Sinha.,Singh,S.K .,2017.Heart Disease Prediction System Using Random Forest.ICACDS,CCIS 721 pp.613-623

[8] Nahar, J., Imam, T., Tickle, K.S., Chen, Y.P.P., 2013. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert Syst. Appl.

40 (1), 96–104.

[9] World Health Organization (WHO), 2017. Cardiovascular diseases (CVDs) – Key Facts. http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-
(cvds).

[10]M,S,Amin.,Y,K,Chiam.,Kasturi Devi,V., 2018.Identification of significant features and data mining techniques in predicting heart disease.36,82-93.

[11] Srinivas, K., Rani, B.K., Govrdhan, A., 2010a. Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng. (IJCSE) 2
(02), 250–255.