Identification of important characteristics and methods for data processing in cardiovascular estimation

[1]Mr.Khodke Harish Eknath, [2] Dr. S. K. Yadav, [3] Dr. D. N. Kyatanavar

[1]Research scholar, [2]Research Guide, [3] Research co-Guide

[1] Dept. of Computer Sci. and Engg.,
Shri JJT University, Chudela,
Jhunjhunu (Rajasthan), India.

Abstract : Cardiovascular disease becomes one among leading risk factors within the global population. One of the most significant problems in the segment of clinical data review is forecasting chronic diseases. In a healthcare sector, the volume of data is massive. The wide variety of raw medical knowledge turns data processing into information to enhance options and estimates. Some early cardiac disease simulation studies have used techniques for data mining. There is also little research which has taken into account the important attributes that forecasting of cardiovascular diseases performs a crucial task. Choosing the best possible mix of key features to use is crucial. It is necessary to pick the appropriate mix of essential aspects. Increase the accuracy of prediction models. The goal of this research is to recognise essential data collection features and techniques to increase the accuracy of cardiovascular forecasting. The building of the prediction models involved a number of combinations of features and seven methods for classification: k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural and Vote networks. Experimental results indicate that a precision of 91.4% in cardiac diseases prediction is obtained by the prediction model for cardiovascular disease that uses the relevant features identified and strongest data analysis. (i.e. voters).

IndexTerms **- Machine learning, Hybrid system, Data Mining, Random Forest, Machine Learning Classifiers,LR,DT,k-NN,SVM.**

I. INTRODUCTION

Death has been the world's number one cause of high blood pressure in recent decades (also referred to as heart disease). The World Health Organisation (WHO) estimated that there were 17.7 million cardiovascular deaths globally in 2015. (WHO, 2017). CVDs are mostly the leading cause of death worldwide: more people die of CVD each year than almost any cause. A few deaths are stopped if coronary disease is expected and noticed in advance.

Applying data mining brings to cardiovascular disease prediction a new dimension. Various techniques of data processing with minimal inputs and efforts are used to identify and extract useful knowledge from the clinical dataset (Srinivas, Rani, & Govrdhan, 2010). Researchers have been investigating different ways to integrate data mining into health care over the last decade to acquire an appropriate diagnosis of cardiac disease.

The success of data processing is largely determined by the methods employed and the features chosen. Patient datasets are redundant and ineffective in the healthcare industry. Data analysis techniques are harder to use without prior and sufficient preparation. The consistency and instability of data in a raw dataset affects Kavitha and Kannan's predicted algorithm results (2016). As a result, efficient training is required in order to prepare the data sets to optimize its capabilities for applying the machine learning algorithms. Moreover, needless properties will lower data extraction technology performance (Paul et al., 2016). For the prediction of cardiac disease with appropriate proper feature-selection techniques and data preparedness, a high degree of precision is appropriate.

Although it has become very evident that the choice of features is as important as the choice of an acceptable method, researchers still fail to adapt proper data mining techniques to the proper set of features. The diagnoses of cardiovascular disease are expected to be extremely accurate, but not easy to receive, Shouman et al. (2013) notes. In addition, it is potentially possible to boost the precision of the calculation by a combination of critical features. This indicates that a systematic experiment to classify important characteristics is needed to serve this aim.

Without the good balance of key characteristics and even the ineffective use of machine learning algorithms, the efficacy of data extraction techniques used in heart failure prediction is dramatically reduced (Dey et al., 2016). Therefore, the optimization process of essential aspects with the highest performing algorithm must be sought. The key purpose of this research is to identify essential elements in the estimation cardiovascular in information analysis methods. However, the proper strategy is not easy to recognize and select. Earlier research tend to show approach in data processing utilized in the forecasting of cardiac problems are unsuccessful and a good study is needed to determine the critical features and methods of data mining that can enhance efficiency. Nahar et al. (2013) notes that it is also critical that the various combinations of features and data mining techniques must be carefully assessed and compared.

There is also a need for rigorous testing to include correct recognition of the information analysis methods and essential characteristics to guarantee that ensure of heart attack is appropriate and precise.

This study specifies the characteristics and strategies of critical information gathering to forecast cardiac diseases. To classify attributes and data mining methods, an experiment was performed. From the UCI Machine Learning Library data base, the heart disease databases were compiled.

The Cleveland data collection is picked because the database is commonly used for the most complete documents by experts in machine learning. Seven classification techniques have been applied using the data set to construct prediction models. Based on

results of the experiment, nine key characteristics and the highest three information analysis strategies were discovered. To confirm for findings of the analysis, the UCI Statlog heart illness dataset was used for analysis.
Furthermore, this review compares the optimum accuracy of the best method found in this study more with greatest precision of the present report.
The substance of the report is coordinated as follows. Discuss several cardiovascular datasets being used describe the key characteristics and strategies of data analysis in this study.
The methods used for the analysis are defined, including data pre-processing, feature selection, classification of data analysis, and quality measures. In order to understand the selection of key characteristics in cardiac disease prediction, the method of dimensionality reduction is developed.
Describes the experiment achievements which are the performance appraisal of the model, which was carried out using seven data analysis methods. To construct the highest model of worth.
The research was conducted to identify core properties and techniques for data analysis. The comparison of the findings using another dataset is listed. Ultimately, in the final section, the results are finalized and the work of the future is defined.

II. DATASETS

The UCI Machine Learning Library has accumulated information about heart disease (Dua and Karra, 2017). There are 04 databases (i.e. Cleveland, Hungary, Switzerland, and the VA Long Beach). For this study, the Cleveland information was selected because it is a popular information with the most detailed records of computer study scientists. The dataset consists of 303 documents. Although the dataset of Cleveland has 76 attributes, the data collection contained in the archive only contains information for a subset of 14 attributes. The Cleveland Clinic Foundation is the archive of the Cleveland data collection. The meaning and type of the attribute are specified in Table 4.1. The predictor of heart disease has 13 features and one feature serves as the performance or the predicted characteristic for a patient's life as cardiac disease. There is an attribute known as 'num' in the Cleveland dataset that displays the incidence of heart attack on varying scales, from 0 to 4. There is a lack of heart dysfunction in this case, 0 as well as all measures between 1 and 4 are people with cardiac failure, where the scale reflects the seriousness of the condition. The representation of the digit function amongst these 303 records is seen in Figure 1.

Cardiovascular disease remains the most deadly disease in today's developed world. This condition affects a person so instantly that he or she barely has time to be handled. The most daunting job for the medical brotherhood is to identify patients accurately on a timely basis. A hospital misdiagnosis leads to a loss of credibility and a bad name. The treatment of this disease is very high at the same time and most of the patients are not available, especially in India. The purpose of this research is to establish cost-effective processing to support the method of supporting database decisions using data mining technology.

About all hospitals use a certain hospital management system to handle patient health care. Unfortunately, the enormous health evidence where critical knowledge is concealed is seldom utilized by most programs. Although these devices produce vast volumes of data in different ways, this data is barely visited and remains untapped. It also needs a great deal of time to make smart decisions. The diagnosis of this condition with several properties or signs is a complex process. This research uses various data processing methods to help in the identification of the disease in question.

Table 1.1: UCI Cleveland cardiovascular data

Sr.No	Attributes Name	Description
01	Age	Patient age
02	sex	Male/female
03	Cp	Chest pain type
04	trestbps	Resting blood pressure
05	chol	Serum cholestoral
06	Fbs	Fasting blood sugar
07	Restecg	Resting ECG results
08	Thalach	Maximum heart rate achieved
09	Old peak	ST depression included by exercise relative
10	Slope	The slope of the peak exercise ST segment
11	ca	Number of major vessels(0-3)

		colored by flouroscopy
12	Thal	3=normal,6=fixed defect,7=reversible defect
13	Expand	Exercise induced angina
14	num	Angiographic disease status

III. METHODS

In the proceedings the UCI Cleveland cardiovascular data gathering was obtained. The data analysis starts with the preparatory step, preceded for functionality, selection of different combinations of attributes and classification models and the creation of forecast templates using information analysis methods. The collection and simulation of functions is replicated with all attribute variations. The loop is iterated as a subset of at least 3 attributes, chosen and added to the structure from the 13 attributes. During each iteration, the efficiency of each framework generated, based on the chosen features and data analysis approach, is reported and the result of the outcome is displayed upon finalization phase.

Details are set out in sections on the pre-processing, function collection, simulation of classifications and efficiency metrics. The effects of strategic objectives indicators are defined in Sections

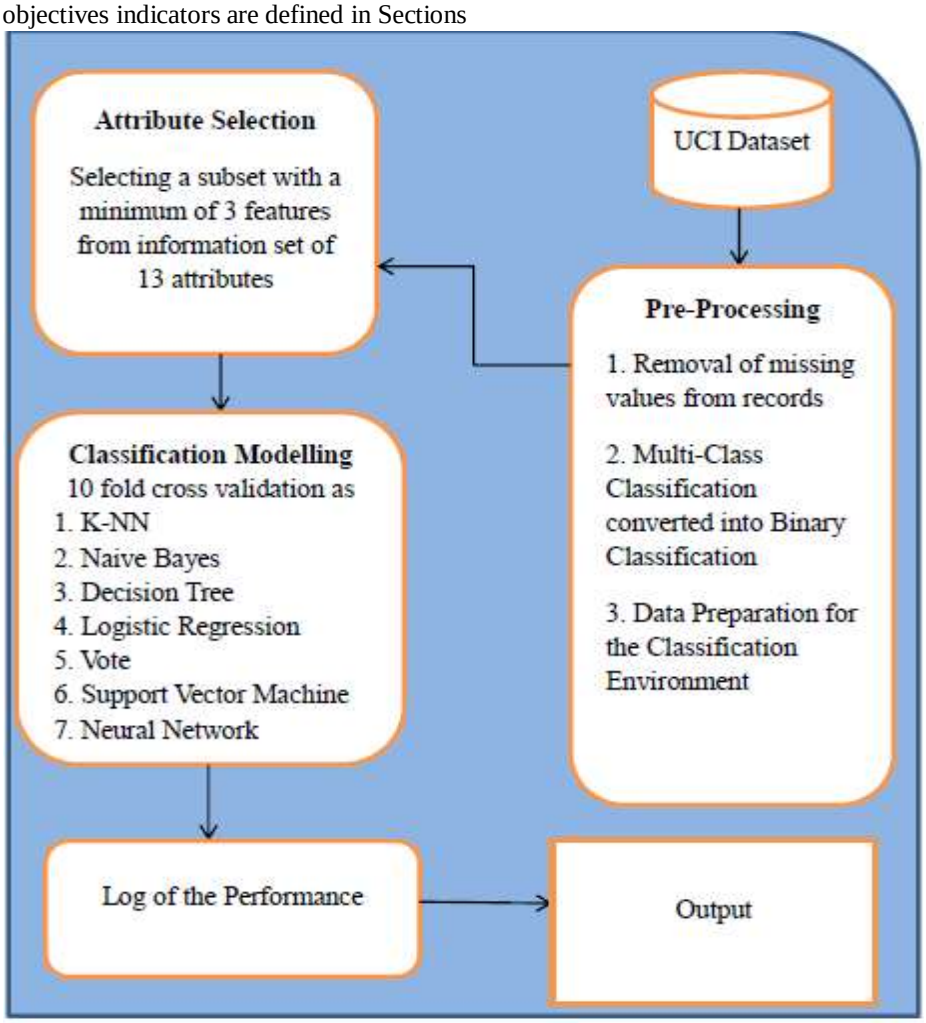

Figure 1.1: Experimental workflow in this analysis

IV. DATA PRE-PROCESSING

The data is pre-processed during treatment. The Cleveland Dataset Values were six missing records. The data set removed all records with missing values and the record number was reduced to 297 from 303. First of all, approximate values for cardiac disease were converted into binary values (0 for elector and 1, 2, three, four for the presence) for the incidence of cardiac disease in the dataset. The task of information extraction was performed by transforming all diagnostic values between 2 and 4 into 1.

Therefore, as the condition value, the resulting dataset involves just 0 and 1, where 0 is the absent and 1 is the effect of hypertension. After reduction and transformation, the distribution of 297 records for 'number' attributes led to 160 records for '0' and 137 records for '1.'

4.1 Feature Selection:

The 13 characteristics shown in the estimation of hypertension refer to the diagnostic records of each case. The rest 11 attributes are all the health characteristics derived from independent screening studies. With 7 classification strategies, the experiment picked a combination of features to create the classification model, such as k-NN, Decision Tree, Naïve Bayes, Logistic Regression, Vote, Vector Assist and the Neural Network. The brute force was then used to restrict its lower restriction (minimum 3 features). Any possible mixture of functions should be checked with both methods. First of all, from the 13 properties of the

experiment, all potential three-charge combinations chosen and tested for each combination using the 7 data analysis methods. The cycle was performed to identify all the possible variations of 4 of the 13 characteristics.
The overall combination usable number by a collection of optimized parameters, except the blank set, should not be less than 3 attributes in this analysis, represented by a 2n- 1A single subset of function mixtures. The subsets of the Combination are also omitted with 2 attributes and 1 attribute. As follows, the total number of variations is determined by the equation.
Combination cumulative number=

$$2^n - \left(\frac{n^2 + n}{2} + 1\right) \qquad (4.1)$$

Where n measures number of characteristics used for the mixture subsets in this experiment that is 13. Therefore 8100 combinations of functions were picked and evaluated in this experiment.

4.2 Data Mining Strategy Classification Simulation

After selecting the features k-NN, Decision Tree, Naive Bavies, Logistic Regressions (LR) and SVM, Neural Network and Voting Techniques (e.g. Naïve Bayes and the hybrid Logistic Regression Technique, these model systems were generated using the seven most traditional classification techniques for data mining. A 10-fold concept of multi approach to check the model's results. This move segments the whole collection of data into 10 sub-sets and analyzes them 10 times. For tests, 9 subsets are included, with the other 1 subset used as an apprenticeship.
Eventually, after adding all 10 implementations, the consequences are shown. Using stratified sampling, the subsets are separated, ensuring each subset has the same class ratio as the primary data set..

4.3 Performance Measure

To determine the efficiency of the classification models, three successful measures have been employed: precision, f-measurement and accuracy. Precision is the proportion of all instances of well-anticipated instances. The Weighted Average Accuracy and Retrieval is the F-measure. For the positive class, accuracy seems to be the proportion of accurate estimates.
This three performance metrics were used to identify essential characteristics, and to evaluate data mining strategies for designing the best modeling accuracy and accuracy. The 3 success metrics offer a clearer explanation including the sum behavior, so that it is easier to grasp essential characteristics. In another hand, research into data analysis methods focuses contributing most powerful mechanisms which can achieve maximum degree of precise diagnosis of cardiac disease, although the most linear estimation quality is consistency and reliability. Output has been separately measured for each classification and for further analysis; all data have been correctly recorded.

IV. RESULTS AND DISCUSSION

The outcomes of prediction performance obtained from the 9 essential attributes and the top three simulation strategies for classification are:
Table 5.1 summarizes the precision of the models achieved through the experiment. This table measures the accuracy of evaluation metrics of selected features and 9 essential attributes..
The forecasting models built on the 9 key characteristics are more accurate than the models developed on all 13 attributes based on Table 4.9. The highest accuracy was the vote for the 13-feature classification model (86.30 percent). In comparison, Vote was also extremely specific in the 9-functional classification model (88.41 percent).

Table 5.1: Accuracy obtained from this study

Accuracy	Vote	Naïve Bayes	Support Vector Machine
Accuracy obtained with 13 attributes	87.30%	85.07%	83.22%
Accuracy obtained with identified 9 significant attributes	**88.41%**	85.81%	86.19%

The findings in Table 5.1 reveal that all of the top three data mining technologies have enhanced the precision of the important characteristics found. This supports the data on the main features of prediction of heart disease. In these, 8 are clinical features obtained from multiple diagnostic tests. The demographic characteristic of the patient is just one, sex. This suggests that clinical studies and monitoring properties have a higher effect than demographic knowledge on the cardiovascular disease detection using data processing methods..
The forecast system which was established employing the Vote and 9 essential attributes of the combination information analysis methodology in Table 4.9 obtained 88.41 percent of full precision.
The second experiment was performed to classify Vote (Vote) as the best efficiency technique, since it superimposes the other two strategies and shows consistencies in both experiments.
The findings prompted more studies to review techniques of combination for information analysis using different information retrieval activities technology combinations to enhance the efficiency of prediction models.

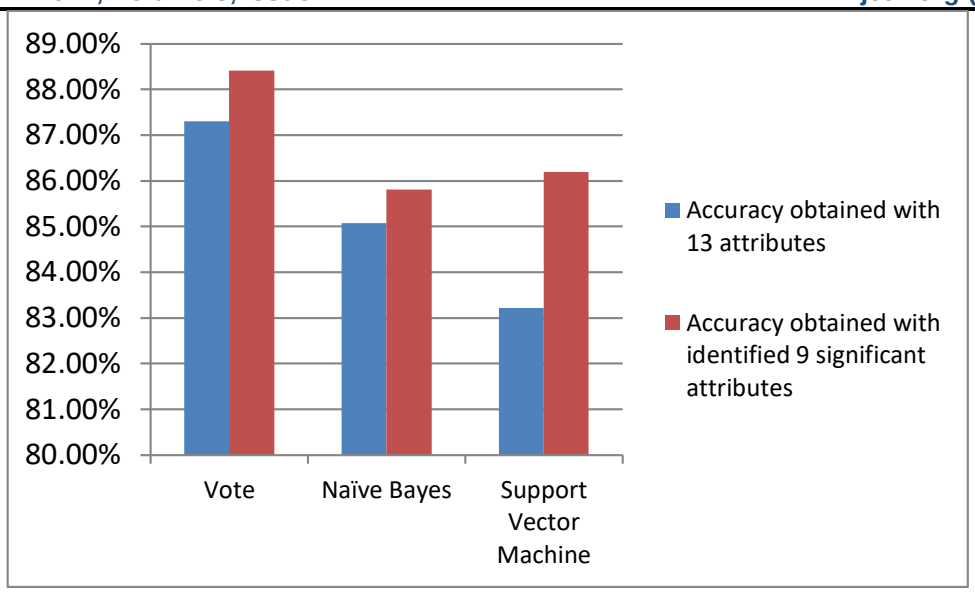

Figure 1.2: Accuracy obtained from this study

In this research, the best prediction model on the basis of evaluation findings were proposed based on the nine essential features and the Vote combination technology. This study shows the overview of the prediction model for theoretical cardiac disease.
This ultimate analysis shows that the usefulness of prediction models has been strengthened by major characteristics and statistical techniques. Such an emerging forecasting model sets the stage for further cardiovascular forecasting trials to support physicians' decision-making in the treatment of heart failure patients.

V.CONCLUSION

The healthcare sector has a large amount of unprocessed patient records. Seeking a way to turn this raw data into a valuable piece of knowledge could save a lot of lives. Data mining methods can be used to process raw data in order to gain new insights and make reliable forecasts for disease prevention. One of the leading causes of death worldwide is heart disease. To avoid death, it is important to diagnose it in patients as soon as possible.
Significant features and the most effective classification modelling techniques that increase the accuracy of heart disease prediction were chosen in this research. To classify the significant features and the top three data mining methods, an experiment was first performed using the UCI Cleveland dataset. Another experiment was conducted using the UCI Statlog dataset to test the results. In this study, nine significant features were chosen.
Vote, Nave Bayes, and Support Vector Machine are the top three data mining techniques that yield high accuracy in prediction, according to this report. The findings of the assessment confirm that the nine features chosen are important.
Furthermore, Vote has outperformed the other two methods in the top three. Using the nine significant attributes and the Vote methodology, the best prediction model was developed. Finally, the proposed model's accuracy was compared to the accuracy of the models proposed in previous research.

REFERENCES

[1] H. E. Khodke, S. K. Yadav and D. N. Kyatanavar, "A new Approach of Heart Disease Prediction system using Data Science," 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), Aurangabad, India, 2020, pp. 225-230, doi: 10.1109/ICSIDEMPC49020.2020.9299627.
[2] H. E. Khodke,"Evaluation for pattern matching and analysis of car's number plate recognition using template matching", INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS, Volume 1, Issue 5 ,Aug 2016,ISSN (Online) 2456-0774.
[3] Khodke Harish Eknath,"study and analysis of medical data mining techniques in healthcare for heart disease using hybrid approach", inflibnet.ac.in, Jhunjhunu,2019.
[4]H.E.Khodke,"pattern matching and analysis of vehicle number plate images for number extraction using template matching and neural network as hybrid method", International Journal Of Current Engineering And Scientific Research,2017.
[5]H.E.Khodke,"pattern matching and analysis of vehicle plate's recognition system using template matching", International Journal for Research in Applied Science & Engineering Technology,volume-4,issue-v.
[6]H.E.Khodke,"pattern matching and analysis of drawn or handwritten digits using correlation",2012, International Journal for Research in Digital Image Processing,vol-4,issue-1.
[7]H.E.Khodke,"pattern matching and analysis of handwritten digits"2012.
[8]Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan,"Identification of significant features and data mining techniques in predicting heart disease",ELSEVIER, Telematics and Informatics 36,2019.
[9]Theresa Princy. R,J. Thomas,"Human Heart Disease Prediction System using Data Mining Techniques",IEEE,2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], ISBN:978-1- 5090-1277-0.
[10]Marjia Sultana, Afrin Haider and Mohammad ShorifUddin,"Analysis of Data Mining Techniques for Heart Disease Prediction",IEEE,iCEEiCT 2016, ISBN:978-1-5090-2906-8.
[11]Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, Khurram Bashir,"Improving Heart Disease Prediction Using Feature Selection Approaches", IEEE,2019,978-1-5386-7729-2.
[12] H. E. Khodke," A smart hybrid system platform for cardiovascular prediction", International Journal of Research and Analytical Reviews,2021