Improving Heart Disease Prediction Using Feature Selection Approaches

[1]Saba Bashir, [2]Zain Sikander Khan, [3]Farhan Hassan Khan, [4]Aitzaz Anjum, [5]Khurram Bashir
[1,2,4,5]Computer Science Department, Federal Urdu University of Arts, Science & Technology, Islamabad, Pakistan
[3]Knowledge & Data Science Research Center, College of Electrical & Mechanical Engineering,
National University of Sciences & Technology (NUST), Islamabad, Pakistan
{[1]saba.bashir3000, [2]zain.sikander50, [3]mrfarhankhan, [4]aitzaz.anjum786, [5]khurrambashir53}@gmail.com
[3]farhan.hassan@ceme.nust.edu.pk

***Abstract*— Heart Disease is the disorder of heart and blood veins. It is very difficult for medical practitioners and doctors to predict accurate about heart disease diagnosis. Data science is one of the more important things in early prediction and solves large data problems now days. This research paper describes the prediction of heart disease in medical field by using data science. As many researches done research related to that problem but the accuracy of prediction is still needed to be improved. So, this research focuses on feature selection techniques and algorithms where multiple heart disease datasets are used for experimentation analysis and to show the accuracy improvement. By using the Rapid miner as tool; Decision Tree, Logistic Regression, Logistic Regression SVM, Naïve Bayes and Random Forest; algorithms are used as feature selection techniques and improvement is shown in the results by showing the accuracy.**

Keywords- Medical data mining, Heart disease prediction, Accuracy, DT (Decision tree), SVM (Support Vector Machine), NB (Naïve Bayes).

I. INTRODUCTION

Heart disease also termed as cardiovascular disease is a major critical condition of the heart and blood veins in majority of deaths. This is the cause of loss because of stroke or heart attack which is 20 percent of all deaths [1]. Different symptoms and causes of heart disease are chest pain, hurt burn and stomach pain, pain in the arms, fatigue and sweating.

A recent study done in 2018 by WHO shows the result that 56.9 million deaths occurred in the world during the year 2016 is due to heart diseases [2]. In 2008 17.3 million people finished because of heart diseases [3]. World health organization recognized the potential of data mining that it can help to predict early stage of heart disease and can provide accurate solution of the disease.

Data mining is basically the discovery of knowledge from huge amount of raw data. Data mining is also known as sub field of data management [4].

Data mining has two main models named as predictive Model and Descriptive Model. Predictive model is defined as a model which is created to predict a particular outcome or result by using predictive modeling techniques [5]. While descriptive model is defined as a model created to provide a better understand of data without targeting a specific variable by using analysis techniques like factor analysis and cluster analysis etc. [6].

Data mining has many learning techniques that can be useful to observe huge pre-existing available data new information. Some examples of the techniques are: Decision Tree (DT), Multi-layer perceptron (MLP), Naïve Bayes (NB), K-nearest neighbor (K-NN) and Support vector machine (SVM) [7].

Many data mining techniques are applied on medical data in order to discover hidden facts from a large amount of data i.e. clustering, regression, classification and outlier etc. Some of the intelligent models in healthcare field are Clinical Support Systems (CSS) and Decision Support Systems (DSS).

Clinical Decision support systems (CDSS) are application of DSS in healthcare field which is designed to support doctors and other health care personnel for improving and making clinical decisions. Decision support systems (DSS) are the information systems used in decision making activities for various fields [8].

Computational intelligence has important role in the prediction of heart disease. Concepts that are used in computation intelligence can discover the relationships between patient attributes and different diseases [9]. In the contemporary studies, many researchers did their work by using feature selection technique in prediction of heart disease. Feature selection is also termed as variable selection or attributes selection. Feature selection is diverse from dimensionality reduction. Feature selection focuses on reducing the number of irrelevant attributes by some techniques i.e attribute subset selection whereas dimensionality reduction reduces the attribute set by generating new attributes from given attribute set [10].

There are many tools in which feature section algorithms and techniques may be applied. This Rapid Miner tool is open source and is easily available on the internet. [11]
.In Future used more dataset and try to provide better result. [12].

A. Motivation

The main motivation of this research is to provide an accurate disease diagnosis framework with reduced feature set. Manually, doctors have to perform number of tests in order to diagnose a particular disease which requires a lot of time, effort and money. Automated disease diagnosis system will

978-1-5386-7729-2/19/$31.00©2019 IEEE

predict the heart disease with high accuracy resulting in time and effort reduction.

B. Research Contribution

Following are some of the major contributions of the proposed research.

- Improving old manual systems
- Prediction of heart disease
- Enhanced accuracy
- Improving efficiency
- Improving effectiveness

II. LITERATURE REVIEW

A recent study done by S. Prakash et al. in 2017 on heart disease prediction which introduced Optimality Criterion feature selection (OCFS) for the extrapolation and proficiently diagnoses the heart disease. Researcher enhances their method for rough set feature selection on information entropy (RFS-IE). In this study they compare the OCFS with RFS-IE in term of computational time, prediction quality, and error rate using different type of data sets. OCFS method can take minimum execution time as compared to the other method [13].

A study is done in 2017 by Seyedamin et al. Researchers works on different machine learning technique and compares their results in term of accuracy. Various machine learning techniques are used in this study on small data set and compared the result with each other. SVM is trained on medical heart disease dataset resulting in a classifier. To improve accuracy aforementioned techniques Bagging, Boosting, Stacking are applied. Using Stacking technique SVM, MLP has best accuracy 84.15% higher than other techniques [14].

In 2015 a research is done by Nguyen Cong Long et al. on disease prediction using firefly algorithm. The classifier is trained by using rough set theory. The results are compared with other classification techniques such as Naïve Bayes and SVM. Proposed work overcomes convergence speed, processing time and increase accuracy to 87.2%. Limitation of this study is that rough set attribute is unmanageable when there is large number of attributes [15].

A study by Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle in 2012 is done which Compares different classifier for the extracting of heart disease. When optimizing absolute precision as a performance measure; SVM has potential. Automated feature selection and motivated feature selection methods are also discussed in this paper such as MFS and CFS. Both techniques show very promising results in term of accuracy [16].

In 2013 Jesmin Nahar et al. used association rule mining classifier, to extrapolate key factor of heart disease. Rule extraction experiment is done on heart disease dataset using rule mining methodology (Aprior,Tertius, Predictive Aprior). Predictive Aprior selects rule based on high accuracy [17].

In 2014 H. Hannah Inbarani et al. performs a research. In this research new administered feature selection methods are imposed for disease prediction which are based on hybridization of Particle Swarm Optimization (PSO) and PSO based Quick Reduct (PSO-QR) .The result of this research shows that on several standard medical datasets there is a surge in proficiency of the proposed technique on the existing feature selection techniques [18].

A research paper "Application of high-dimensional feature selection: evaluation for genomic prediction in man" is published in 2015 in which conclusion of five different feature selection techniques is scrutinized on the basis of performance effect of (G-BLUP) and (Bayes C) methods. This study predicts high density lipoprotein cholesterol (HDL) and body mass index (BMI). Result of this study shows that supervised feature selection of SNP in the (G-BLUP) granted flexible and computationally alternative to Bayes C. The limitation in this study is that when supervised selection is used then predictive performance requires so much careful evaluation otherwise results may not be achieved [19].

SinaTabakhi published a paper in Elsevier with the topic of "An unsubstantiated feature selection algorithm based on ant colony optimization". In this study feature selection is classified as unsubstantiated, filter and multivariate. The method used in this research trade –off between computational time and value of results. Experimental results show increase in efficiency and effectiveness of the UFSACO method and also show improvement over previous related methods [20].

A research paper with the topic of "Classification of heart disease using artificial neural network and feature subset selection" is published by M. Akhil Jabbar et al. This paper research shows that feature subset selection is a method which is used to reduce dimensionality and input data. This paper introduces a classification technique by using ANN and feature selection for classification of heart disease. By reducing the number of elements, the number of diagnosis tests which are needed by doctors from patient are also reduced. Andhra pardesh data set is used in this research and results show that accuracy is enhanced over the outdated classification techniques. Moreover the results also show that this system is faster and precise [21].

A survey paper is published by Divia Tomar and Sonali Agarwal in which educes the importance of various data mining techniques i.e. classification, clustering, association and regression etc. in the study of health. They also reveal introduction of these techniques, their advantages and disadvantages. They also highlighted hindrances and further problems regarding data mining techniques on health care data. This paper is recommended as a suit able choice to study the available data mining techniques [22].

Table 1 shows the comparison of different research papers.

TABLE 1: COMPARISON OF DIFFERENT DATA MINING TECHNIQUES

Ref	Year	Technique	Type	Data set	Accuracy
[1]	2017	Machine Learning	OCFS	UCI	82.3%
[2]	2017	Machine Learning	MLP/ SVM	UCI	84.15%
[3]	2013	Machine Learning	MFS/ CFS	Cleveland	83.2%
[4]	2013	Rule mining	Predictive	UCI	83.65%

III. Proposed Approach

The proposed approach used to complete this research is started by downloading an open source UCI data set. After verifying the dataset, next step is preprocessing and data discretization in the form of Data cleaning, Data Transformation, Data Reduction, Binning and Select Attributes. After applying all these techniques on downloaded dataset, the main technique feature selection is applied. Later on, following algorithms are applied on the data i.e Decision Tree, Logistic regression, Logistic regression SVM, Naïve Bayes and Random forest. After applying algorithms and techniques we compare results and discuss about conclusion. The flow of these techniques is shown in Fig 1.

A. *UCI Dataset*

UCI dataset is an open online source which is associated with multitudes of diseases and covers a large source of databases, domain theories and data generators which are utilized by the researchers.

B. *Preprocessing & discretization*

Preprocessing of data is presented in an intelligible presentation by turning raw data into fathomable context for a purpose.

C. *Data Cleaning*

Data cleaning is a process in which data is cleaned by removing missing data, duplicate data and resolving data inconsistencies. As a result data quality is improved resulting in usefulness of data.

.

D. *Data Transformation*

Conversion of data or information from one format to another format is known as data transformation. It usually done when a source format is needed to convert into required format for a specific purpose.

E. *Data reduction*

Transformation of numeric or alphabetic digital information into a corrected ordered and simplified form experimentally or empirically. The main concept of data reduction is to reduce multitudinous amounts of data into useful information.

F. *Binning*

Binning divides the groups and number of continuous values in to small bins by using equal frequency or equal depth binning techniques.

G. *Feature Selection:*

Feature selection is also denoted as variable selection, Attribute selection or variable subset selection for model construction which inhibits the process of choosing a subset of pertinent features (variable predictors).

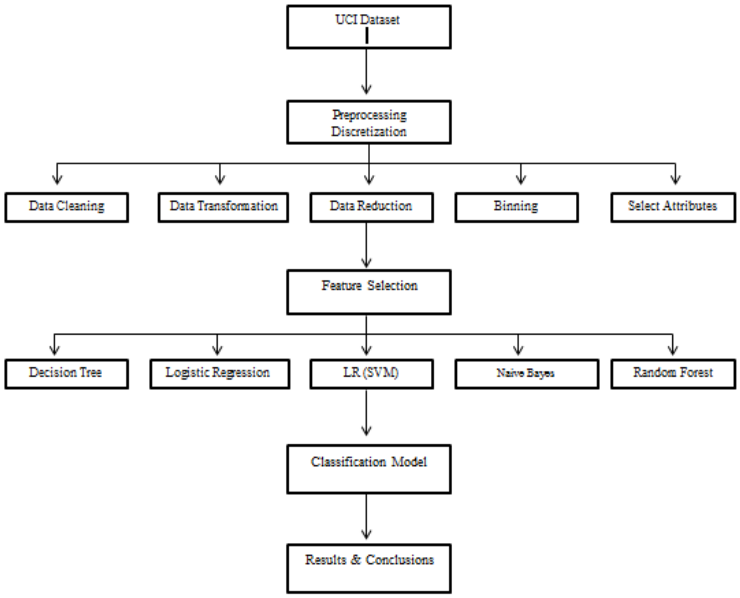

Fig 1: Flowchart of proposed approach

H. *Classification Algorithms*

Following classification algorithms [23-26] are then applied on the preprocessed dataset.

- *Decision Tree:* A decision support which is a decision tree implements a tree like graph and possible outcomes which includes chance event outcomes, resource codes and purpose. It is a representation of an algorithm that consists of conditional control statement.
- *Logistic Regression:* Logistic regression is assessing the parameters of logistic model in regression analysis.
- *Logistic Regression SVM:* Logistic regression SVM is more precise and generates better results as compared to SVM. Medical data is used to generate the model and to show results.
- *Naïve Bayes:* Naïve Bayes classifiers perform classification by calculating the probability of given dataset. Each attribute in given data is considered as independent of other. High probability class is the output class of given instance.
- *Random Forest:* Random forest is a tree based method that is used for both classification and regression analysis. Multiple trees are constructed and the mean prediction would be the output for classification.

I. *Results and conclusions*

In this phase results are presented and compared with previous outcomes of the researches. If there are any enhancements then the final interpretation will be presented and the final decision will be based on the classifier which has high accuracy on the given dataset.

IV. Experiments, Results & Discussion

The application of techniques reveal the results of all five applying algorithms Decision Tree, Logistic regression, Logistic regression SVM, Naïve Bayes and Random forest by using 5-fold cross validation that are presented in this section.

A. *Preparation of Data*

In very first step a medical data set has been downloaded from an open source database named UCI repository. In next, irrelevant rows are excluded according to specific requirement of heart disease as other researchers done in literature review i.e. total rows or attributes are more than 300 we reduced or select 14 attributes that are most relevant for the prediction of heart disease i.e age, sex, status etc.

B. *Tool*

Rapid Miner is a data analysis tool which already has instilled techniques to be implemented on the model and prepared data then we convert data from csv file format to xlsx file format then import this data into the rapid miner in order to obtain the desired results.

C. *Implementation*

Then apply ensemble process on data with sub process Minimum Redundancy Maximum Relevance Feature Selection (MRMR). The output after applying (MRMR) feature selection is in the form of weights. The next step is Data validation which is the main step of this experiment. We apply all five techniques as a sub process on the weighted data and results are obtained.

Experimental Results

- *Decision Tree*

After applying the Decision we achieve accuracy 82.22 % which doesn`t surpasses the previous researcher`s results.

- *Logistic Regression*

After applying the Logistic Regression accuracy is 82.56% which is also not as per desired achievement.

- *Random Forest*

After applying the Random Forest we achieve quite an improvement in accuracy which 84.17% but still less than logistic regression SVM and Naïve Bayes.

- *Naïve Bayes*

After applying the Naive Bayes, increase has been achieved in accuracy which is 84.24 %. This increase is also more than previous researches.

- *Logistic Regression SVM*

After applying the logistic Regression (SVM), increase has been shown in accuracy which has highest accuracy as compared to other classifiers and is 84.85% which meets our research goal.
The basic structure of methodology is shown in Fig 2 which describes how different techniques are applied on given dataset.

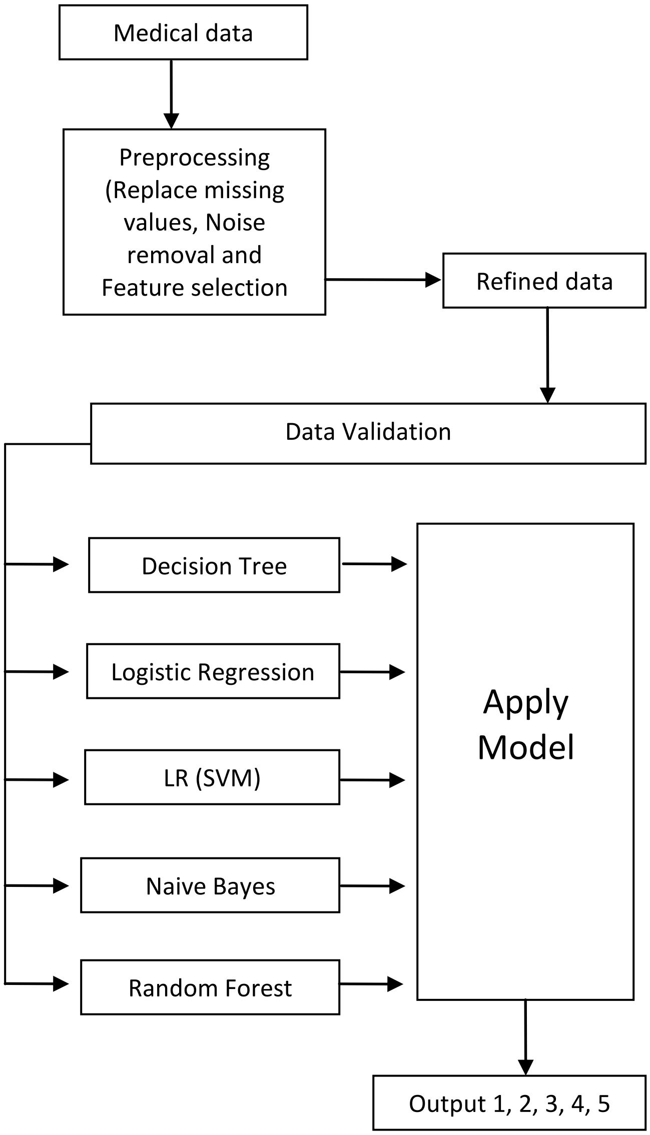

Fig 2: Proposed methodology structure

Table 2 shows the comparison of data mining techniques that are applied on UCI heart disease dataset.

TABLE 2: COMPARISON OF ACCURACY RESULTS

Sr.No	Technique	Technique	Data Set	Accuracy
1	MRMR/FS	Decision Tree	UCI	82.22%
2	MRMR/FS	Logistic Regression	UCI	82.56%
3	MRMR/FS	Random Forest	UCI	84.17%
4	MRMR/FS	Naïve Bayes	UCI	84.24%
5	MRMR/FS	Logistic Regression (SVM)	UCI	84.85%

V. CONCLUSION & FUURE WORK

The main ambition of this paper is to improve accuracy in prediction of heart disease by using feature selection techniques. Different data mining techniques i.e. Decision Tree, Logistic regression, Logistic regression SVM, Naïve

Bayes and Random forest are applied individually in Rapid miner on a UCI heart disease date set and compared results with the past researches. This research achieves the goal which was as per expectation and accuracy has been improved from previous mentioned values in literature review. As shown in Table 2, the accuracy of Decision Tree is 82.22%, Logistic Regression 82.56%, Random Forest 84.17%, and Naïve Bayes 84.24% and Logistic Regression SVM is 84.85.

As result shows clearly that increase in accuracy has been achieved in two of techniques Logistic Regression (SVM) and Naïve Bayes applied in this paper.

As if we compare this increase with the previous used techniques then the highest accuracy achieved was 84. 15 %. Random forest, Naïve Bayes and Logistic Regression (SVM) all three techniques increase accuracy than previously implemented machine learning techniques. If we compare these techniques with rule mining which is also used previously with achievement of 83.60 % accuracy, the proposed research improved much higher accuracy than rule mining.

As Logistic regression is with higher accuracy achievement so this paper suggests Logistic Regression as a best feature selection technique for predicting heart disease.

In future these techniques can also be applied on real time medical datasets and also can be used in the form of ensembles i.e combinations of multiple techniques. This would result in increase of further accuracy and high performance.

REFERENCES

[1] World Congress of Cardiology Scientific Sessions 2016 Volume 11, Issue 2, Supplement, Pages e1-e203, June 2016

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

[3] ARCHANA, BADE, AHER DIPALI, and SMITA KULKARNI PROF. "International Journal On Recent and Innovation Trends In Computing and Communication." 2277-4804.

[4] Silwattananusarn, Tipawan, and Kulthida Tuamsuk. "Data mining and its applications for knowledge management: A literature review from 2007 to 2012." arXiv preprint arXiv:1210.2872 (2012)

[5] Leventhal, Barry. "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.

[6] Leventhal, Barry. "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.

[7] 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017

[8] McKhann, Guy, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. "Clinical diagnosis of Alzheimer's disease Report of the NINCDS- ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." Neurology 34, no. 7 (1984): 939-939.

[9] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40, no. 4 (2013): 1086-1093.

[10] El Mountassir, Mahjoub, Slah Yaacoubi, José Ragot, Gilles Mourot, and Didier Maquin. "Feature selection techniques for identifying the most relevant damage indices in SHM using Guided Waves." In 8th European Workshop On Structural Health Monitoring, EWSHM 2016. 2016.

[11] Prakash, S., K. Sangeetha, and N. Ramkumar. "An optimal criterion feature selection method for prediction and effective analysis of heart disease." Cluster Computing (2018): 1-7.

[12] Pouriyeh, Seyedamin, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, and Juan Gutierrez. "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease." In Computers and Communications (ISCC), 2017 IEEE Symposium on, pp. 204-207. IEEE, 2017.

[13] Long, Nguyen Cong, Phayung Meesad, and Herwig Unger. "A highly accurate firefly based algorithm for heart disease prediction." Expert Systems with Applications 42, no. 21 (2015): 8221-8231.

[14] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach." Expert Systems with Applications 40, no. 1 (2013): 96-104.

[15] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40, no. 4 (2013): 1086-1093.

[16] Inbarani, H. Hannah, Ahmad Taher Azar, and G. Jothi. "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis." Computer methods and programs in biomedicine 113, no. 1 (2014): 175-185.

[17] Bermingham, Mairead L., Ricardo Pong-Wong, Athina Spiliopoulou, Caroline Hayward, Igor Rudan, Harry Campbell, Alan F. Wright et al. "Application of high-dimensional feature selection: evaluation for genomic prediction in man." Scientific reports 5 (2015): 10312.

[18] Tabakhi, Sina, Parham Moradi, and Fardin Akhlaghian. "An unsupervised feature selection algorithm based on ant colony optimization." Engineering Applications of Artificial Intelligence32 (2014): 112-123.

[19] Jabbar, M. Akhil, B. L. Deekshatulu, and Priti Chandra. "Classification of heart disease using artificial neural network and feature subset selection." Global Journal of Computer Science and Technology Neural & Artificial Intelligence 13, no. 3 (2013).

[20] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5, no. 5 (2013): 241-266.

[21] Hand, David J. "Principles of data mining." Drug safety 30, no. 7 (2007): 621-622.

[22] Roiger, Richard J. Data mining: a tutorial-based primer. Chapman and Hall/CRC, 2017.

[23] Kavakiotis, Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal15 (2017): 104-116.

[24] Goswami, Saptarsi, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti, and Basabi Chakraborty. "A review on application of data mining techniques to combat natural disasters." Ain Shams Engineering Journal 9, no. 3 (2018): 365-378.