An Analysis of Heart Disease Prediction using Different Data Mining Techniques

Nidhi Bhatla
GNDEC, Ludhiana, India

Kiran Jyoti
GNDEC, Ludhiana, India

ABSTRACT
Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the abnormal health conditions that directly influence the heart and all its parts. Heart disease is a major health problem in today's time. This paper aims at analyzing the various data mining techniques introduced in recent years for heart disease prediction. The observations reveal that Neural networks with 15 attributes has outperformed over all other data mining techniques. Another conclusion from the analysis is that decision tree has also shown good accuracy with the help of genetic algorithm and feature subset selection.

Keywords
Heart disease; Data mining; Fuzzy logic; Decision tree; Naive bayes; Classification via clustering; Neural networks; Weka tool; Genetic algorithm.

1. INTRODUCTION
Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Healthcare industry today generates large amount of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden patterns and also provides healthcare professionals an additional source of knowledge for making decisions. Figure 1 depicts the basic data mining process model.

The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure – a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease also includes functional problems of the heart such as heart-valve abnormalities or irregular heart rhythms. These problems can lead to heart failure, arrhythmias and a host of other problems.

Effective and efficient automated heart disease prediction systems can be beneficial in healthcare sector for heart disease prediction. Our work attempts to present the detailed study about the different data mining techniques which can be deployed in these automated systems. This automation will also reduce the number of tests to be taken by a patient. Hence, it will not only save cost but also the time of both, analysts and patients.

2. METHODOLOGY
This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts or practitioners for accurate heart disease diagnosis. The main methodology used for our work was by examining the publications, journals and reviews in the field of computer science and engineering, data mining and cardiovascular disease in recent times [5].

3. RESEARCH OBSERVATIONS

3.1 Data Mining and Neural Networks
An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system. In this work, Heart disease prediction system has been developed using 15 attributes [4]. Earlier 13 attributes were used for prediction but this research work incorporated 2 more attributes, i.e. obesity and smoking for efficient diagnosis of heart disease.

The data mining tool Weka 3.6.6 is used for experiment. Initially, missing values were identified in the dataset and they were replaced with appropriate values using ReplaceMissingValues filter from 3.6.6 [4]. Further, various data mining techniques have been analyzed on heart disease database. Confusion matrix is obtained for each classifier.

Table 1 depicts the outcomes of this research work and it shows that neural networks has outplayed over other data mining techniques.

Classification Techniques	Accuracy
Naive Bayes	**90.74%**
Decision Trees	**99.62%**
Neural Networks	**100%**

Table 1: Comparison of various data mining techniques

3.2 *Fuzzy Logic and Genetic Algorithm*

The proposed method in this research work is an extended version of the model that combines the genetic algorithms for feature selection and fuzzy expert system for effective classification. Fuzzy set theory and fuzzy logic are highly suitable for developing knowledge based systems in healthcare for diagnosis of diseases [2].

Experiments are conducted in Matlab using fuzzy tool. For this, Mamdani model of fuzzy system is used. The fuzzy rules are generated based on experts' knowledge in this domain. The dataset from UCI machine learning repository is used, and only 6 attributes are found to be effective and necessary for heart disease prediction. In the proposed system, the input is the set of all the selected features and the output of the system is to achieve a value 0 or 1 that indicates the absence or presence of heart disease in patients.

In fuzzy logic process, initially fuzzification is performed by collecting the crisp set of input data and converting it to a fuzzy set using fuzzy linguistic variables, fuzzy linguistic terms and membership functions. After that, an inference is made based on a set of rules and lastly, defuzzification step is performed [2]. This system generates the fuzzy rules based on the support sets obtained. Table 2 shows this support set.

S.No.	Attributes	Support Set	
		Heart Patients	Non – Heart Patients
1.	Chest Pain Type	4	1,2,3
2.	Rbps	134-153	142-154
3.	Exang	Yes	No
4.	Oldpeak	2.06-6.2	<2.06
5.	Thalach	71-136	136-168
6.	Ca	1,2,3	0

Table 2: Values of the features in the support set

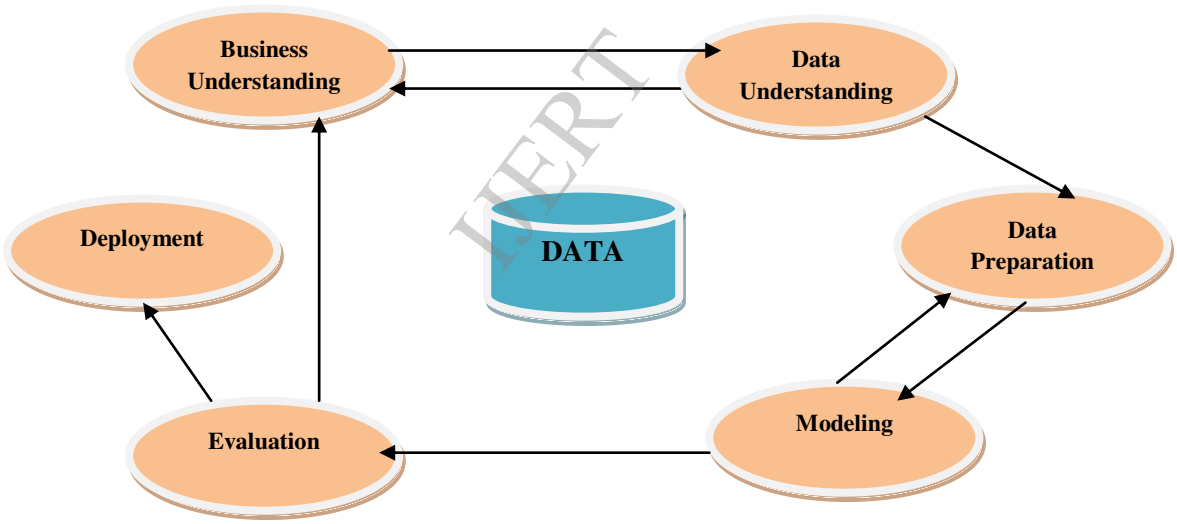

Figure 1: Data Mining Process Model

3.3 *Data Mining and Supervised Machine Learning Algorithms*

This research work has presented the data classification based on various supervised machine learning algorithms, namely, Naive Bayes, Decision List and KNN. TANAGRA tool is used to classify the data and the data is evaluated using 10-fold cross validation.

TANAGRA [20] is a data mining tool for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. It provides an easy-to-use interface by allowing the users to analyze either real or synthetic data. This tool also proposed an architecture to the users allowing them to easily add their own data mining methods, to compare their performances. It is a wide set of data sources, direct access to data warehouses and databases, data cleansing, interactive utilization.

Experiments are conducted by using the training data set of 3000 instances with 14 different attributes. Depending upon the attributes, the dataset is classified into two parts, i.e. 70% of the data is used for training and rest 30% is used for testing. Performance of each algorithm is determined and comparison is made based on the accuracy and evaluation time of calculation for each algorithm [12]. It has been observed that Naive Bayes algorithm performed better in comparison to

other two algorithms. Table 3 illustrates the performance study of various algorithms.

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

Table 3: Performance analysis of various Algorithms

3.4 Data Mining and Genetic Algorithm

The objective of this work was to reduce the number of attributes which were used for heart disease diagnosis. Earlier, 13 attributes were used for this prediction but this research work reduced the number of attributes to six only using Genetic Algorithm and Feature Subset Selection.

Genetic Algorithm [6] incorporates natural evolution methodology. The genetic search started with zero attributes, and an initial population with randomly generated rules. Based on the idea of survival of the fittest, new population was constructed to match with fittest rules in the current population, as well as offspring of these rules. Offspring were generated by applying genetic operators; cross over and mutation. The process of generation continued until it evolved a population P where every rule in P satisfied the fitness threshold. With initial population of 20 instances, generation continued till the twentieth generation with cross over probability of 0.6 and mutation probability of 0.033. The genetic search resulted in six attributes out of thirteen attributes.

CFS Evaluator is also used in addition to the genetic algorithm. Observations are conducted using Weka 3.6.0 tool. Initially, data set of 909 records with 13 attributes was used. All attributes were made categorical and inconsistencies were resolved for simplicity. After reduction of 13 attributes to 6 attributes, various classifiers are used on the dataset corresponding to these 6 attributes for heart disease prediction. Performance analysis of these classifiers is shown in Table 4. It can be perceived from the table that Decision Tree has outperformed with highest accuracy and least mean absolute error.

DM Techniques	Accuracy	Model Construction Time	Mean Absolute Error
Naive Bayes	96.5%	0.02s	0.044
Decision Tree	99.2%	0.09s	0.00016
Classification via Clustering	88.3%	0.06s	0.117

Table 4: Comparison Table for three Classifiers

3.5 IHDPS and Data Mining Techniques

This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Networks. IHDPS is web-based, user-friendly, scalable, reliable and expandable system which is implemented on the .NET platform [15].

IHDPS can discover and extract hidden knowledge associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus help healthcare analysts and practitioners to make intelligent clinical decisions which traditional decision support systems cannot. It also helps in reducing treatment costs by providing effective treatments. Moreover, it displays the results both in tabular and graphical forms. This IHDPS is based on 15 attributes.

A total of 909 records were obtained from the Cleveland Heart Disease database. The records were equally divided into two datasets, i.e. training dataset (455 records) and testing dataset (454 records). It has been observed during the analysis that Naive Bayes appears to be most effective as it has the highest percentage of correct predictions (86.53%) for patients with heart disease, followed by Neural Network (85.53%) and Decision Trees. Decision Trees, however, appears to be most effective in case of predicting patients with no heart disease, i.e. (89%) as compared to other two models.

DM Techniques	Accuracy
Naive Bayes	86.53%
Decision Trees	89%
ANN	85.53%

Table 5: Performance analysis of IHDPS

4. RESULTS

For better understanding, results for each data mining techniques have been shown separately in different tables. Various classifiers are employed in combination with different data mining techniques for heart disease prediction. It can be perceived from the observations that in some cases, the same classifier produces different accuracy for different data mining techniques.

5. CONCLUSIONS

The objective of our work is to provide a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. The analysis shows that Neural Network with 15 attributes has shown the highest accuracy i.e. 100% so far. On the other hand, Decision Tree has also performed well with 99.62% accuracy by using 15 attributes. Moreover, in combination with Genetic Algorithm and 6 attributes, Decision Tree has shown 99.2% efficiency.

6. REFERENCES

[1] P .K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules"; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40.

[2] E.P.Ephzibah, Dr. V. Sundarapandian, "Framing Fuzzy Rules using Support Sets for Effective Heart Disease Diagnosis"; International Journal of Fuzzy Logic Systems (IJFLS) Vol.2, No.1, February 2012.

[3] A.Sudha, P.Gayathri, N.Jaisankar, "Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability"; International Journal of Computer Applications (0975 – 8887) Volume 41– No.17, March 2012.

[4] 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.

[5] Jyoti Soni, Sunita Soni et al., "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.

[6] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm"; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[7] E.Sivasankar, Dr.R.S.Rajesh, "Knowledge Discovery in Medical Datasets Using a Fuzzy Logic rule based Classifier"; 978-1-4244-7406-6/10/$26.00, IEEE, 2010.

[8] M.A. Saleem Durai, et. al. "Effective analysis and diagnosis of lung cancer using fuzzy rules"; International Journal of Engineering Science and Technology Vol. 2(6), 2102-2108, 2010.

[9] Mostafa Fathi Ganji, Mohammad Saniee Abadeh, "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease"; Proceedings of ICEE 2010, May 11-13, 2010, 978-1-4244-6760-0/10/$26.00©2010 IEEE.

[10] Huang Hai, "Data Mining Based on a Compensative Fuzzy Neural Network"; International Conference On Computer Design And Applications (ICCDA), 2010.

[11] M.A.Saleem Durai, N.Ch.S.N.Iyengar, "Effective Analysis and Diagnosis of Lung Cancer Using Fuzzy Rules"; International Journal of Engineering Science and Technology, Vol. 2(6), 2010.

[12] Asha Rajkumar, G. Sophia Reena, "Diagnosis of Heart Disease Using Datamining Algorithm"; Global Journal of Computer Science and Technology, Page | 38 Vol. 10 Issue 10 Ver. 1.0 September, 2010.

[13] Shantakumar B.Patil, Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network"; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.

[14] Rupa G. Mehta, Dipti P. Rana, Mukesh A. Zaveri, "A Novel Fuzzy Based Classification for Data Mining using Fuzzy Discretization"; World Congress on Computer Science and Information Engineering, 2009.

[15] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques"; 978-1-4244-1968-5/08/$25.00©2008 IEEE.

[16] Cleveland database: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[17] Han, J., Kamber, M, "Data Mining Concepts and Techniques"; Morgan Kaufmann Publishers, 2006.

[18] American Heart Association. Heart Disease and Stroke Statistics — 2004 Update. Dallas, Tex.: American Heart Association; 2003.

[19] Statlog database: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart

[20] http://eric.univ-lyon2.fr/~ricco/tanagra/