e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTION OF HEART DISEASE USING DATA MINING TECHNIQUES

A.Nikila*1, S. Sabari*2

*1,2Department of Computer Science, Velalar College Of Engineering And Technology,Erode, India.

ABSTRACT

Many variables influence a person's heart throughout daily life. Many problems are arising at a quick rate, and new heart disorders are being discovered at an alarming rate. In today's world of stress, the heart is an important organ in the human body that pumps blood throughout the body for proper blood circulation and a healthy lifestyle. The primary goal of this study is to present a heart disease prediction model for predicting the occurrence of heart disease. Furthermore, the goal of this research is to find the best classification system for detecting cardiac disease. Medical practitioners face a challenging task in detecting the likelihood of heart disease in a patient because it requires years of experience and extensive medical testing. Three data mining classification techniques are addressed and used to construct a prediction system in order to determine and predict the risk of heart disease in this work KNN classification, SVM classification, Nave Bayes, and Random Forest. The major goal of this important research project is to find the best classification algorithm for achieving maximum accuracy in the classification of normal and abnormal people. As a result, it is feasible to prevent the loss of lives at an earlier stage. When compared to other algorithms, the Random Forest algorithm performs better. The project is designed using R Language 3.4.4 with R Studio.

I. INTRODUCTION

A human heart's influencing factors. Many problems are arising at a quick rate, and new heart disorders are being discovered at an alarming rate. Heart, as a key organ in the human body that pumps blood through the body for blood circulation, is essential in today's world of stress, and its health must be preserved for a healthy lifestyle. A person's life events define the health of their heart, which is fully based on professional and personal activities. A form of heart disease may also be handed down through generations due to many hereditary factors. According to the World Health Organization, more than 12 million people die each year due to various causes around the world. A human heart's affecting factors. Many problems are arising at a quick rate, and new heart disorders are being discovered at an alarming rate. Heart, as a key organ in the human body that pumps blood through the body for blood circulation, is essential in today's world of stress, and its health must be preserved for a healthy lifestyle. The health of a human heart is decided by a person's life experiences and is entirely dependent on professional and personal activities. A form of heart disease may also be handed down through generations due to many hereditary factors. According to the World Health Organization, more than 12 million people die each year due to various causes around the world. The symptoms of heart disease are very dependent on the type of discomfort experienced. Some symptoms are difficult for the average person to recognize Bad eating habits, lack of sleep, restless nature, depression, and a variety of other factors including such rotundity, poor diet, family history, high blood pressure, high cholesterol levels, idle geste, family medical history, smoking, and hypertension may all contribute to an increased risk of heart disease in young people. The assessment of heart diseases is extremely vital, and it is also the most difficult work in the medical world.

When the croaker asses and understands the instances through homemade check-ups at regular intervals, all of the above criteria are taken into account. The symptoms of a cardiac issue are very dependent on the type of discomfort experienced by a person. Some symptoms are not commonly associated by the general public. Casket discomfort, dyspnea, and cardiac pulsations are all common symptoms. Angina, or angina pectoris, is a type of heart pain that develops when a portion of the heart doesn't get enough oxygen. Angina can be brought on by stressful circumstances or physical activity, and it usually lasts less than 10 seconds.

Heart attacks can also occur as a result of other heart conditions. The symptoms of a heart attack are similar to those of angina, except that they can occur at any time and are usually more severe. Heart attack symptoms can occasionally mimic dyspepsia.

A heavy feeling in the casket, as well as heartburn and stomach pains, can cause this. Other signs of a cardiac event include pain that spreads throughout the body, such as from the head to the arms, neck, back, stomach, or

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

jaw, dizziness and flightiness, gushing sweats, nausea, and nausea. Breathlessness can occur as a result of heart failure, which occurs when the heart becomes too weak to circulate blood.

Some cardiac problems have no symptoms at all, particularly in older adults and people with diabetes. Sweating, high levels of weariness, quick twinkle and breathing, shortness and chest pain are all indications of a natural heart problem. Still, these symptoms may not appear until a person has reached the age of 13.

In these situations, forming an opinion becomes a complex undertaking requiring extensive knowledge and talent. A threat of a heart attack or the potential of a heart complaint, if linked ahead of time, can aid cases in taking preventative and nonsupervisory actions. Recently, the healthcare industry has been accumulating massive amounts of data on instances, and their complaint opinion reports are being used to analyse heart attacks all across the world. Machine literacy methods can be used to analyse large amounts of data about cardiac complaints.

Data mining is the process of extracting critical decision-making data from a database of old records for future study or vaticination. Without data mining, the information could be hidden and unidentifiable. The categorization is a data mining technique that allows for the creation of unborn offspring or prognostications based on the actual data provided. Medical data mining allowed for the integration of bracket techniques and training on the dataset, which led to the exploration of retired patterns in medical data sets that were utilised to analyse the case's unborn condition. As a result, medical data booby-trapping can provide insight into a case's history and can be used to provide clinical support through analysis. These patterns are extremely important in the clinical examination of cases. In layman's terms, medical data mining employs bracket algorithms, which are critical for determining the likelihood of occurrence.

To make prognostications that determine the person's nature of being impacted by heart complaint, classification algorithms can be trained and evaluated. The supervised machine learning idea is used in this research to make predictions. Prognostications are made using a comparative comparison of three data mining bracket algorithms: Random Forest, Decision Tree, and Nave Bayes. The analysis is carried out separately in numerous cross-confirmation settings and several chance split evaluation styles.

In this study, the dataset from the UCI machine learning repository was used to make cardiac complaint prognostications. When the cardiac complaint dataset is used for training, the prognostications are made utilising the bracket model that is created using the bracket algorithms. This final model can be used to examine various cardiac diseases.

II. LITERATURE REVIEW

Every day, sophisticated computer-based systems collect vast volumes of data in the healthcare industry using automatic data record systems, from which data mining can extract useful knowledge. The next section briefly discusses cardiac disease and how data mining techniques can be used to treat it.

RELATED WORK

The authors of this research (1) indicated that the prognosis for patients with heart failure remains bleak. The goal of this study was to estimate the most essential factors for prognosticating patient survival and to describe instances to estimate their survival chances utilising data mining styles, as well as the most appropriate health-care fashion. Five hundred and thirty-three cases of cardiac arrest were included in the study.

They used Bayesian networks to perform traditional statistical analysis and data mining analyses. RESULTS The 533 patients had an average age of 63 (17), with 390 (73) men and 143 (27) women making up the sample. Cardiac arrest was seen in 411 (77) cases at home, 62 (12) cases at a public place, and 60 (11) cases on a public route. The belief network of the variables revealed that age, copulation, the original heart measurement, the cause of heart failure, and the technical reanimation methods used are all directly related to the chance of being alive after heart failure.

Physicians could use data body-trapping tactics to forecast case survival and alter their methods accordingly. This work could be done for any medical treatment or problem, and data from a service or a croaker may be used to swiftly generate a decision tree. The contribution of the data mining system in classifying variables and swiftly determining the relevance or impact of the values that a variable takes on the study's purpose was highlighted when traditional analysis was compared to data mining analysis. The system's key constraint is acquiring knowledge, as well as gathering enough data to develop an usable model.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Cardiac arrest is characterised as an unrecoverable robotic general rotational arrest caused by cardiac inefficiency. It is identified when there is no femoral user feels for more than 5 seconds. Cardiac arrest without reanimation results in sudden cardiac death. Because the survival rate for cardiac arrest cases is believed to be between 1 and 20, the public health impact of unanticipated cardiac mortality is significant. This corresponds to fatalities in the United States and deaths in France at the same time. The case profile is already well-known, as it primarily affects men aged 45 to 75.

The hospitalisation process must be efficient and effective. The technique for supposing blame varies according on the type of cardiac arrest, and some studies demonstrate that colourful ways are preferred over others depending on the reason of cardiac arrest. According to the American Heart Association, around one million people in the United States will have a coronary attack each year.

Ninety-five percent of unintentional cardiac arrest patients die before reaching the hospital, and heart disease kills more lives than the next six leading causes of death combined (cancer, habitual lower respiratory conditions, accidents, diabetes mellitus, influenza and pneumonia). The majority of people in the United States who die from heart disease are under the age of 65. These findings demonstrate the interest in predicting the risk of death following heart failure, and there is a need to analyze the events that occurred during care in order to provide prognostic information.

Classic statistical analysis have already been conducted and have provided some information on the epidemiology and causes of heart failure. The use of a probability in a statistical technique to outline heart failure in a sample of patients and prognosticate the influence of some events in the care process is presented in this study.

They came to the conclusion that the Bayesian network could be effective in medical analysis for exploring data and discovering hidden links between events or sample characteristics. This is a first stage in agitating healthcare professional conflicts. The requirement for enough data to discover chronicity in the links is the fundamental limitation of these techniques.

The authors of this publication noted that the spadework in this subject has been done after a decade of abecedarian interdisciplinary exploration in machine literacy; the 1990s should witness widespread use of knowledge discovery as a tool for creating knowledge bases. The implied benefits of this exploration aroused the contributors to the AAAI Press book Knowledge Discovery in Databases. The editors hope that some of this enthusiasm will be reflected in AI Magazine compilations of this work.

It is projected that every three months, the world's quantum of information doubles. Database size and number are likely rapidly increasing. The world's total number of databases was expected to reach five million in 1989, however the majority of them were tiny DBASE III databases. Because even ordinary transactions, like as a phone conversation, the usage of a credit card, or a medical test, are typically logged in a computer, the robotization of business activities creates an ever-growing flood.

Academic and governmental databases are rapidly expanding as well. Previously, the National Aeronautics and Space Administration had far more data than it could analyse. Remote sensing data satellites, which are expected to launch in the 1990s, are estimated to contribute one terabyte (1015 bytes) of data per day, more than all previous activities. To look through the filmland produced in one day at a pace of one picture per alternate, a person would have to labour numerous evenings and weekends. The government financed Human Genome Project will store hundreds of bytes for each of the billions of inheritable bases in biology.

A million million bytes of 1990U.S. story data produce patterns that depict the cultures and mores of today's United States in a similar manner. What should we do with all of this raw data? It's likely that only a small portion of it will ever be seen to mortal eyes.

If it is to be understood at all, it will have to be anatomized by computers. Simple statistical methods for data analysis have long been developed, but advanced methods for intelligent data analysis have yet to emerge.

As a result, the gap between data generation and data comprehension is widening. Simultaneously, there is a rising acceptance and expectation that data, when appropriately anatomized and presented, will be a valuable resource that can be exploited to gain a competitive advantage. The computer wisdom group is responding to the scientific and practical challenges of locating information adrift in a sea of data.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Michie (1990), a top European specialist on machine literacy, predicted that "the next area that's going to explode is the utilization of machine literacy tools as an element of large-scale data analysis," when predicting the future of AI. Data mining was identified as one of the most potential exploration motifs for the 1990s in a recent National Science Foundation workshop on the future of database exploration.

Some exploration styles have been well-developed enough to be included in commercially accessible software. ID3 is used in some expert system shells to transform rules from exemplifications. Inductive, neural net, and inheritable literacy approaches are used by other systems to find patterns in specific computer databases. Many forward-thinking firms are sifting through their databases for interesting and helpful patterns using these and other methods.

American Airlines scans its frequent flyer database for its best customers, focusing its marketing efforts on them. Farm Journal analyses its subscriber statistics and employs innovative printing technology to create hundreds of editions tailored to specific demographics. Several banks have derived superior loan blessing and ruin vaticination techniques using patterns observed in loan and credit histories. General Motors is deciding individual expert systems for colourful models based on a database of machine issue reports. Manufacturers of packaged goods are scouring grocery scanner data for information on goods' elevations and purchase trends.

Increasing requests for, as well as greater effort to provide, tools and approaches for database discovery have resulted from a combination of corporate and exploratory interests. This is the first book on the subject to bring together cutting-edge research from around the world. Inductive literacy, Bayesian statistics, semantic query optimization, knowledge access for expert systems, information proposition, and fuzzy sets are only a few of the ways to discovery covered.

The book is meant for everyone who are interested in or engaged in computer wisdom and data operations, and it will both enlighten and inspire further research and operations. Professionals working with databases and operating information systems, as well as those applying machine literacy to real-world issues. The present state of knowledge discovery in databases is quite diverse. We'll start by defining and describing the terminologies that will be used throughout this investigation.

The nontrivial creation of hidden, previously unknown, and possibly beneficial information from data is known as knowledge discovery. We describe a pattern as a statement S in L that describes relations among a set FS of F with a confidence c, comparable to how S is easier (in some ways) than reciting all data in FS, given a collection of data (data) F, a language L, and some measure of certainty.

Knowledge is a pattern that is both attractive (according to a stoner's interest scale) and certain (according to the stoner's standards). Discovered knowledge is the result of a software that watches a database's set of data and generates patterns. These definitions of language, certainty, and simplicity and interestingness measurements are intentionally imprecise in order to accommodate a wide range of methods. These concepts, taken together, represent our understanding of the abecedarian aspects of database discovery. The meanings of these concepts are shown in the following paragraphs, and their relevance to the problem of knowledge discovery in databases is suggested.

They pointed out that Jonathan Swift's Gulliver faced one of the most important discovery processes during his visit to the Academy. "Where they plant three or four Words that might compose Part of a Judgment, they mandated them to Scribes," the "Design for developing academic Knowledge by practical and mechanical procedures" was producing sequences of words by arbitrary variations. Although this approach has the potential to create a large number of interesting judgements in the long term, it is currently hampered.

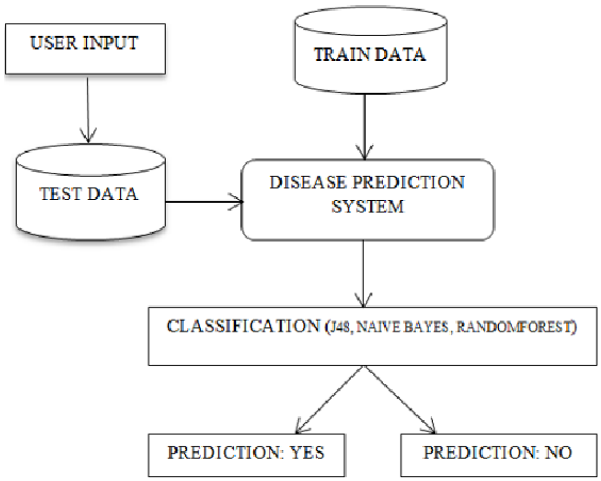
III. APPLICATIONS OF DATA MINING

Scholars of data mining have long investigated the use of tools and equipment to improve data processing in large and complicated datasets. In the field of medicine, data mining techniques are critical for diagnosing, predicting, and comprehending healthcare data. Treatment center analysis aimed at refining treatment policies and preventing any errors in hospitals, early disease diagnosis, disease prevention, and hospital death reduction are some of these uses.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

DATASET COLLECTION

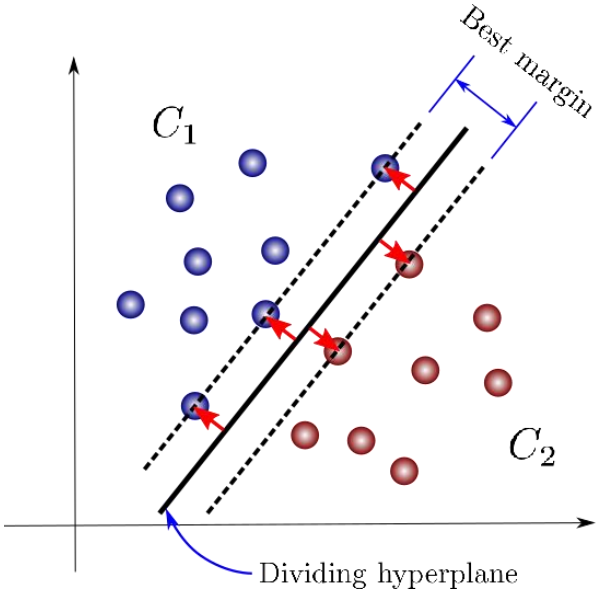
This module takes the heart dataset, which includes features such as age, gender, kind of chest discomfort, resting blood pressure, serum cholesterol fasting blood sugar level and disease



SVM CLASSIFICATION

It is a regression and classification analysis method based on machine learning. Learning algorithms are used to train controlled learning models. They analyze a big amount of information to find trends. By constructing two parallel lines, an SVM generates parallel divisions. In a high-dimensional space, almost all attributes are used for each category of data. It divides the space into flat and linear segments in a single pass. A distinct divide between the two categories should be as large as possible. Use a plane called a hyperplane to partition the data. To divide provided data into classes, use the biggest margin in a high-dimensional space.

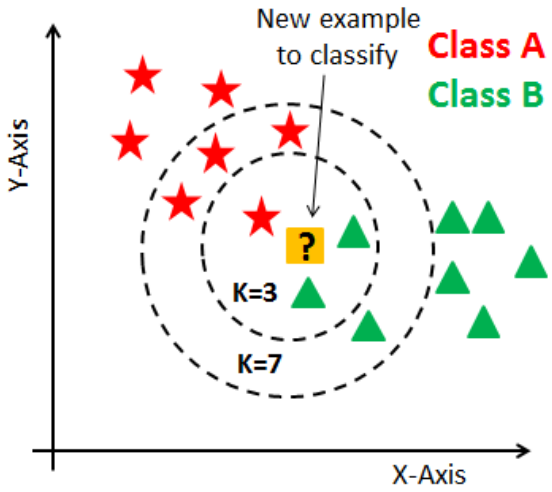
The distance between the two classes' closest data points is represented by the margin between them. The lower the classifier's generalisation error, the larger the margin. In this module, the records are categorised as disease 1 or 2 using SVM classification



KNN CLASSIFICATION

This module uses KNN classification with a K value of 6 and a disease column as the binary classification factor. 75% of the data is given as training data and 25% as testing data. The disease kind and the record number for the testing data are discovered and displayed as a result.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com



NAIVE BAYES CLASSIFICATION

The Naive Bayes classifier is based on Bayes' theorem and predictor independence assumptions. A Naive Bayesian model is simple to construct and does not require iterative parameter estimation, making it ideal for huge datasets. Despite its simplicity, the Naive Bayesian classifier often works admirably and is commonly used because it outperforms more complex classification algorithms. From P(c), P(x), and P(x|c), the Bayes theorem can be used to get the posterior probability, P(c|x). The effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors, according to the Naive Bayes classifier. Class conditional independence seems to be the term for this assumption

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

RANDOM FOREST

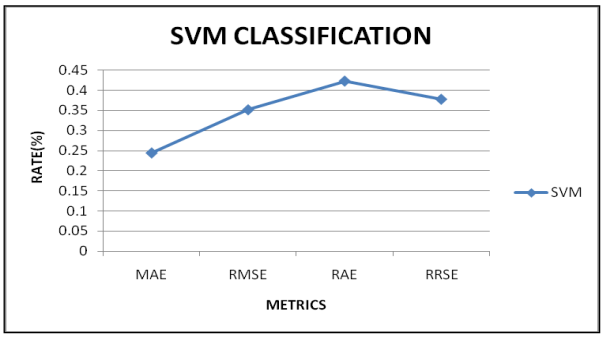
Random forests (RF) are a decision tree-based mixture of tree predictors in which each tree is dependent on the values of a random vector collected separately and uniformly across the forest. The strength of individual trees in the forest and their association determine the generalisation error of a forest of tree classifiers. In terms of noise, they are more resilient. It is a supervised classification technique that is used for prediction and is regarded superior than decision trees because of the high number of trees in the forest. Typically, the trees are trained independently and their forecasts are averaged. Based on the problem domain, the random forest approach can be used for both classification and regression.

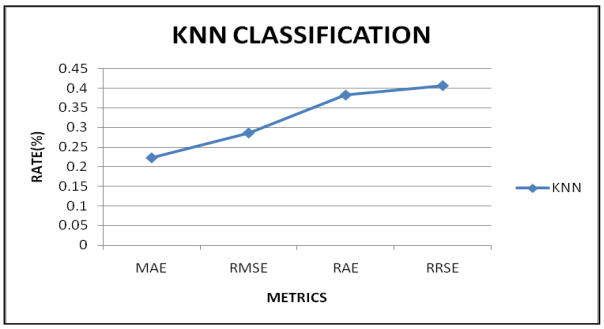
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

IV. PERFORMANCES METRICS ANALYSIS

The Heart Complaint Analysis Model's evaluation criteria are shown below. The following is a list of the items in the table. SVM, KNN, RF, and Naive Bayes. Delicacy values are used to analyze the performance of the above stated algorithms.

METRICS	SVM	KNN	NB	RF
MAE	0.245	0.223	0. 237	0.217
RMSE	0.352	0.2 86	0.332	0.282
RAE	0.423	0.3 82	0.402	0.285
RRSE	0.378	0.406	0.398	0.302
Accuracy	0.891	0.908	0.894	0.952





V. CONCLUSION

The proposed strategy generates a better understanding of cardiac complaint prediction using new data mining methods such as SVM, RF, NB, MLP, and DT the weighted association classifier. The DT bracket is a popular way to group the qualities from a case record. The bracket performance and delicacy of the heart complaint opinion can be improved by SVM clustering and DT with weighted association classifier. Experts and professional croakers of heart specialists have backed up all of these criteria. The proposed fashion is more effective than the current system. In future study, we will seek more breakthroughs to improve the accuracy of cardiac complaint prediction by including criteria and croakers suggestions inside various sorts of medical terms, as well as providing first-aid advice.

ACKNOWLEDGEMENT

We would like to e Express our gratitude towards our Project guide Dr.S.Vivekanandan(Assistant professor,VCET) for his constant guidance, supervision and constructive criticism during the successful completion of the research work.

VI. REFERENCES

[1] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in Health Technology and Informatics, Vol. 107, No. 2, pp. 1256-1259, 2004.

[2] W.J. Frawley and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview", AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.

[3] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Intelligent Computing in Signal Processing and Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[4] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[5] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008

[6] Shantakumar B. Patil and Y.S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol. 31, No. 4, pp. 642-656, 2009.

[7] Nidhi Singh and Divakar Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", Ph.D Dissertation, Department of Computer Science and Engineering, Barkatullah University Institute of Technology, 2012.

[8] Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: "Tapping the Power of Text Mining", Communication of the ACM. 49(9), 77-82, 2006.

[9] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006

[10] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998

[11] A comparison of antiarrhythmic-drug therapy with implant-able defibrillators in patients resuscitated from near-fatalventricular arrhythmias. The Antiarrhythmics versusImplantable Defibrillators (AVID) Investigators. N Engl JMed. 1997 Nov 27;337(22):1576-83.

[12] R.Wu, W.Peters, M.W.Morgan, "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal of Healthcare Information Management. 16(4), pp. 50-55, 2002.

[13] Mary K.Obenshain, "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, vol. 25, no.8, pp. 690–695, Aug. 2004.

[14] G.Camps-Valls, L.Gomez-Chova, J.Calpe-Maravilla, J.D.MartinGuerrero, E.Soria-Olivas, L.Alonso-Chorda, J.Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery." Trans. Geosci. Rem. Sens. vol.42, no.7, pp.1530–1542, July.2004.

[15] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[16] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management. 16(4), 50-55, 2002.

[17] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[18] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases",http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.

[19] Mohd, H., Mohamed, S. H. S.: "Acceptance Model of Electronic Medical Record", Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005.