A Method for Prediction of Heart Disease Using Support Vector Machine (SVM) Algorithm

Priyadarshni Kumari[1], Chinmay Bhatt[2], Varsha Namdeo[3]
CSE Department, SRK University, Bhopal, India[1, 2, 3]

Abstract- *Data mining is a technique widely employed in the medical industry to retrieve disease-related data. Heart disease is the most common human ailment. The big data created during cardiac disease prediction is too complicated and large to analyze and evaluate using existing approaches. Using data mining methodology reduces the time required to accurately anticipate illness. To diagnose and extract meaningful information from the large dataset, this thesis presents a hybrid of SVM and KNN. To solve the problem of cardiac disease, the simulation program MATLAB is employed. Coronary infection is a vicious infection that affects millions of individuals worldwide. Early detection is vital given the high loss of life charges and large number of people who suffer from cardiovascular sickness. Studying a disease isn't always easy. Using a recovery investigation form to an ai for coronary disease gives more security than mining human administration data. DB mining today, clinical test findings are frequently decided based on specialists' insight and expertise rather than the rich data available in several vast databases. This method frequently causes unintentional predispositions, lapses, and high medical costs, affecting the quality of patient care. Many doctors' offices now use secure data management systems to store patient and social security information. These data frameworks often produce a lot of data in different formats, such as statistics, for clinical decision making. The heart is an essential organ. This extensive database is occasionally used by the heart to operate efficiently. If the heart is not working properly, it will affect the brain, kidneys, etc. It is a pump that circulates blood throughout the body. Inefficient blood circulation harms organs like the brain, and when the heart stops pumping, death happens within minutes. Heart efficiency is vital to life. Heart disease is a disease of the heart and its blood vessels. Several variables raise the risk of heart disease. These methods and strategies are used to find knowledge from databases. Some data mining methods describe.*

Keywords:- Data Mining, Heart disease, SVM, algorithms, brain suffer, Decision Trees MATLAB.

Introduction

In today's culture, a larger proportion of the population has an impact on heart disease. In this rapidly growing society, people are required to maintain a luxurious standard of living. They operate as if they were machines, with the goal of acquiring a substantial sum of money while maintaining a comfortable lifestyle. There is no way for them to focus on the organs of their body. The number of risk factors for heart disease is increasing on a daily basis. Depending on their employment conditions, people alter their dietary practices. It raises blood pressure and sugar levels in those of a young age. Individuals consult with a medical professional on a more frequent basis than they do not. They are taking a medicine that is unique to them. Symptoms are exacerbated by these types of approaches. Each year, a disproportionately large number of people become aware of the dangers of cardiovascular disease. One of the most significant causes of gloom and mortality is heart disease. According to the World Health Organization, heart disease is responsible for 12 million deaths worldwide each year, with risk factors such as smoking and high blood pressure able to be reduced by 50%. [9]. According to current estimates, heart illnesses are estimated to be the main cause of 35 to 60% of all deaths expected worldwide by 2025. [1]

A high percentage of false positives continues to plague the present algorithms for disease diagnosis, which is related to the unstructured nature of the information. When a dataset is insufficient or partial, it can result in false alarms in the diagnosis of accurate outcomes. There is a lack of concentration in the existing work on data preprocessing, which results in a significant stumbling block in the total process. As an alternative to using raw data, the future work of this survey will begin with a data preparation method to construct a complete dataset instead of starting with raw data. Data Mining Challenges in the Healthcare Industry Data mining in health care is confronted with challenges from a variety of perspectives. It develops into a complicated procedure for the purposes of assortment, retrieval, and analysis. It must be input and kept in a systematic manner in order to be used in future applications. It is necessary to design a standardization programmer in this manner if the standard is to be followed consistently.

Data mining is a collection of computational approaches that are used to uncover instructional patterns from large amounts of raw data. The healthcare industry currently generates large volumes of diverse data, including information on hospital resources, computerized patient records, illness diagnosis, and other topics. This is critical. A massive volume of data must be generated and evaluated for knowledge removal in order to provide effective assistance for comprehending the current situations in the healthcare business. Among the data mining operations are the following: formulating a hypothesis, collecting data, conducting preprocessing, guesstimating the model, interpreting the model, and visualizing the findings [1]. Preliminary considerations should be given to understanding the many types of data mining algorithms available and how they operate before proceeding to the study of how these algorithms are applied to medical data.

It came into being somewhere around the middle of the 1990s and quickly established itself as a powerful tool for extracting necessary information from large amounts of data. In general, Knowledge Discovery (KDD) and Data Mining are related terms that are often used interchangeably. However, some researchers believe that the two terms are not interchangeable because Data Mining is one of the majorities of critical stages in the Knowledge Discovery (KDD) process, which is incorrect. According to Fayyad et al., the knowledge discovery process in a database is systematized into four stages: the first stage is data selection, in which data is gathered from disparate sources; the second stage is preprocessing the selected data; the third stage is data transformation, in which the data is transformed into an appropriate format so that it can be processed supplementary; the fourth stage is data mining, in which a suitable data mining method is applied to the transformed data for extrapolation; and the final stage is knowledge discovery.

II. Related Work

Heart disease is a leading cause of death in today's society, and numerous diagnostic tools have been created by a number of researchers to help identify the condition. It is the purpose of this section of the dissertation to present previous work accomplished in the field of heart disease in the health care business, which is described as follows:

Lamia Abed Noor Muhammed [2] presented and conferred on the experiment that was carried out using the naive bayes method in order to build a predictive model as an artificial diagnose for heart disease based on a dataset that contained a restricted set of parameters that had previously been measured for individual participants. Evaluation of the outcomes with different approaches should be carried out following the use of the related data that was provided from the UCI repository data.

Patients who approach for hospitalisation for heart disease have accumulated medical information, and algorithms are being used to prolong that information and outcome will be provided in a variety of easily understandable words and graphs, according to Ranganatha S., Pooja Raj H.R., Anusha C., and Vinay SK [3]. Due to the presence of extraordinarily large data sets, data mining algorithms (here referred to as ID3 and Nave Bayesian algorithms) are employed. A decision tree is generated by ID3 and can be easily comprehended by the user. The naive Bayesian approach predicts the likelihood of heart disease based on the information provided.

G.Vaishali and V.Kalaivani [4] used big data to create a centralised serene monitoring system that they implemented. In the proposed system, a massive collection of medical records will be used as an input. It is intended to extract the necessary information from the testimony of heart patients from this medical dataset using the map reduce technique. Heart infection is the most serious health problem, and it is also one of the major causes of death all over the world. Aiming for earlier identification of cardiac disease has emerged as a critical concern in the medical research community. Some characteristics of heart disease exposure, such as the RR interval, QRS interval, and QT interval, are investigated. As part of the classification process, the classification phase determines if the patient is normal or abnormal. The discovery stage uses the map reduce technique to detect disease and minimise the dataset. As a result, the proposed approach contributes to the classification of a large and complicated medical dataset as well as the perception of cardiac disease.

Using a previously created heart illness database record, Ankita Dewan and Meghna Sharma [5] constructed a prototype that can discover and extract previously undiscovered knowledge (patterns and relationships) connected to heart disease. It is capable of resolving intricate inquiries for the detection of heart disease and, as a result, may assist medical practitioners in making elegant clinical decisions, something that conventional decision support systems were not capable of. It is possible to lower the cost of therapy by delivering high-quality care at an affordable price.

In this paper, B. Venkatalakshmi and M.V Shivsankar [6] presented and extended a predictive mining-based diagnosis and forecast method for cardiac disorders. Different experiments have been carried out in order to measure the performance of distinct predictive data mining methodologies, which include Decision Trees and Naive Bayes algorithms, in the same dataset. As a resource data, a 13 attribute ordered medical database from the University of California, Irvine Machine Learning Repository has been used in this projected study. The decision tree and the Naive Bayes methods have been applied, and their performance in terms of identification has been evaluated. When compared against the Decision Tree, the Nave Bayes algorithm outperforms it.

For the purpose of establishing neural network weights, S. U. Amin and colleagues [7] developed a hybrid system that makes use of the global optimization help of a genetic algorithm for the purpose of establishing neural network weights. Predictions of heart disease are based on risk factors such as family history, age of onset, diabetes, high cholesterol, tobacco use, hypertension, alcohol consumption, and obesity.

A. K. Sen, S. B. Patel, and D. P. Shukla [8] proposed a layered neuro-fuzzy technique to predict the occurrence of coronary heart disease that was modelled in the MATLAB simulation tool. While performing an investigation for the occurrence of coronary heart disease, the execution of a neuro-fuzzy integrated technique provided an extremely low error rate and a high level of job effectiveness.

M. Jabbar, P. Chandra, and B. Deekshatulu [9] proposed a unique approach for association rule mining (ARM) based on sequence number and clustering transactional data set for heart disease prognosis using a transactional data set based on sequence number and clustering transactional data set. As a result of this, the proposed approach may be regarded scalable and efficient. The implementation of the proposed approach was implemented in the C programming language and condensed primary memory required by considering a minuscule cluster at a time.

The authors [10] developed class association rules utilising feature subset assortment to forecast a model for heart disease. P. Chandra, M. Jabbar, and B. Deekshatulu were involved in the research. The association rule draws conclusions about relationships between attributes such as ethics, and the classification rule predicts the class in the patient dataset. As with genetic search, which defines traits and contributes to the prediction of cardiac problems, feature selection analyses how well a feature performs.

Mai Shouman, Tim Turner, and Rob Stocker [11] hypothesised a number of solitary and hybrid data mining techniques for use in the diagnosis of cardiovascular illness. The use of a single data mining method for the detection of cardiac disease has been extensively explored, with the results demonstrating high levels of precision. The results of recent research have proven that combining more than one diagnostic procedure will result in improved results in the diagnosis process. They identified a gap in the diagnosis of heart disease and the treatment required for it, and they proposed a model to close those gaps in order to determine whether applying amalgam and solitary data mining techniques in heart disease treatment can result in reliable performance of data mining techniques. Dissimilar data mining techniques such as Nave Bayes, Decision trees, Neural Networks, Multilayer Perceptrons, and Kernel density can be applied to dissimilar heart disease datasets and the precision of each technique can be measured in this article. Following that, using hybrid data mining approaches to various heart disease datasets reveals that the accuracy of the results varies. The results of a comparison of both strategies in the diagnosis of heart disease using the Cleveland heart disease datasets revealed that hybrid techniques had greater accuracy and precision than individual data mining techniques.

Using Naive Bayes classification, Dhanashree S. Madhekar, Mayur P. Bote, and Shruti D. Deshmukh [12] proposed a classifier approach for heart disease prediction and shown how it may be utilised for this purpose. The proposed system will organise medical data into five distinct categories: no, low, average, high, and extremely high. No, low, average, high, and extremely high are the first four categories. In addition, the system will consider the class label of dissimilar unknown samples, if any, and for this forecast, the two fundamental functions explicitly classification (training) and prediction (testing) will be carried out. The precision of the system will be determined by the application of distinct algorithms and techniques to dissimilar databases. The study is carried out on Cleveland datasets, with the classification algorithms being implemented in order to forecast group membership for data instances. Training and testing on datasets will be carried out in the classification approach. The training datasets are provided as input to the classifier, and the classified dataset is used in conjunction with the testing function to perform the testing function. System will classify the threat of heart disease into three categories: low, average, and high. The system will classify the threat of heart disease into three categories based on the Naive Bayes method and then test the dataset.

III. Proposed Work

Heart disease dataset [17] and pre-process dataset by converting into xlxs format are used in this section. Identifying the Entropy and information gain of each attribute independently is how the Entropy-based feature assortment will be accomplished. Using the decision tree technique, it will be possible to calculate the measure of information gain. Entropy is a measure of impurity that is often employed. When impurity levels are higher, the information content is higher. In artificial intelligence, Information Gain is a performance parameter that indicates how well one attribute AI identifies the training information. The Decision Tree separates the training group in a recursive manner. The entropy parameter is the one that best classifies data (H). Entropy is a useful metric for assessing the amount of information carried by a collection of events. The entropy of the set S is indicated by the letter H. (S). If S=Sample of n training events and Pi is the chance of an event occurring, then the entropy can be calculated as follows:

$$H(S) = \sum i=1\ n - \ Pi \log 2Pi$$

Calculate the information gain for each attribute separately. Data classification accuracy is a statistical metric that measures how well an attribute categorises the data. We have calculated the information gain (Gain(S, Ai)) for each attribute using the Algorithmic Approach, and the attribute with the highest information gain will be selected for decision-making at the end of the process. Sv is the subset of S for which the value v is assigned to element A..

$$Gain(S, Ai) = H(S) - \sum v\epsilon Values(Ai)\ P(Ai=v)\ H(Sv)$$

It is our measure of how well one attribute Ai categorizes the training data, and it is derived from information gain.

When the process is normalized, it provides linear transformation on the original range of data, which is achieved by utilizing the min-max algorithm. This method specifically fits the data by generating a new range from an existing one range, which is unique. In this paper, we use a novel technique to level 2 classifications, in which SVM is used for level 1 classification and KNN is used for level 2 classifications. Support vector Machine classification is used to begin the level 1 classification process on both the training and test data sets. Following the prediction, predicted classes will now be passed for training. It is planned to use KNN-based Level 2 classification on the training data.. It will be calculated the classified outcome of the new classes goal0, goal1, goal2, goal3, and so on. Each of the four classes' performance and forecast will be determined by computing standard metrics such as sensitivity, accuracy, and specificity. All of these parameters are investigated and compared to the parameters of the preceding SVM and KNN-based algorithms.

IV. Proposed Model

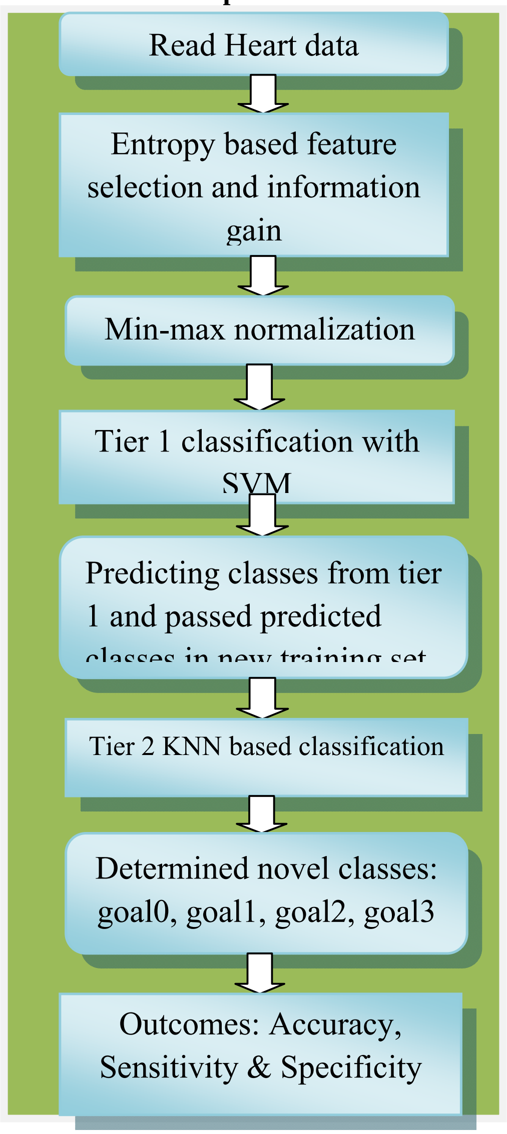
Read Heart data
Entropy based feature selection and information gain
Min-max normalization
Tier 1 classification with SVM
Predicting classes from tier 1 and passed predicted classes in new training set
Tier 2 KNN based classification
Determined novel classes: goal0, goal1, goal2, goal3
Outcomes: Accuracy, Sensitivity & Specificity

Fig. 1: Diagram of proposed method.

V. Result Analysis

Here, we provide the results of our thorough tests to compare the performance of M.KNN, SVM, and the suggested 2-tier technique on real health-care data from a Chinese city, which were conducted in the previous section. In order to conduct the experiments, we use the MATLAB platform, which is comprised of three Intel 3.4 GHz workstations with each having 16GB of RAM.

There are 76 attributes restricted in this database, but all published trials refer to employing a subset of 14 of those attributes. According to diligent researchers, the Cleveland database is the only one that has been utilized by ML researchers to date. The "end" field is used to refer to the presence of heart disease in the peaceful. From 0 (there is no such thing) to 4, it has integer values. Experiments with the Cleveland record have been focused on attempting to distinguish between presence (values 1, 2, 3, 4) and absence (values 1, 2, 3, 4) of the record (value 0). [17] There are only 14 attributes that are used.

This record has 76 qualities, with the exception that all published studies refer to a subset of 14 of them being used. According to the most thorough researchers, the Cleveland record is the only one that has been used by ML researchers up to this point. [38] The presence of cardiac disease in the patient is indicated by the "end" field in the form of a tick mark. It has integer values ranging from zero (no presence) to four. Experiments with the Cleveland record have been focused on attempting to distinguish between presence (values 1, 2, 3, 4) and absence (values 0 and 1). (value 0).

The result analysis of the proposed work is performed using the accuracy and specificity parameter and for the processed attributes confusion matrix is formed which is exposed below in table 1, table 2 and table 3:

Table 1: Confusion Matrix for SVM

Confusion Matrix for SVM	
Confusion Matrix of 'c2'	
98	2
35	16
0	0
Confusion Matrix of 'c1'	
60	15
64	12
0	0
Confusion Matrix of 'c3'	
95	4
39	13

0	0
Confusion Matrix of 'c4'	
136	4
9	2
1	0

Table 2: Confusion Matrix for Proposed Methodology

Confusion Matrix for Proposed Methodology	
Confusion Matrix of 'c2'	
246	21
14	22
0	0
Confusion Matrix of 'c1'	
233	15
22	33
0	0
Confusion Matrix of 'c3'	
251	17
14	21
0	0
Confusion Matrix of 'c4'	
281	9
6	7
0	0

Table 3: Confusion Matrix for Proposed 2tier

Confusion Matrix for Proposed 2tier	
Confusion Matrix of 'c2'	
246	9
14	20
0	0
Confusion Matrix of 'c1'	
233	27
22	40
0	0
Confusion Matrix of 'c3'	
251	9
14	19
0	0
Confusion Matrix of 'c4'	
281	3
6	7
0	0

5.4.1 Accuracy Analysis

The capacity of a test to correctly distinguish between patients and healthy instances is referred to as its accuracy. We should determine the proportion of true positives and true negatives in all of the analyzed cases in order to estimate the accuracy of a given test. This can be expressed mathematically as follows:
$\text{Accuracy} = (TP+TN)/TP+TN+FP+FN$
When looking at this parameter, the researchers conducted a comparison between SVM, M.KNN, and the 2tier projected method, and they discovered that the accuracy rate of SVM is approximately 75 percent, that of M.KNN is 89 percent, and that of our method is approximately 90 percent, indicating that our method generates a higher accuracy rate than the existing SVM method.

Table 4: Accuracy result analysis of the projected 2tier method

	Accuracy		
	SVM	M.KNN	2-tier
goal2	0.754967	0.88488	0.9174
goal1	0.476821	0.877888	0.861386
goal3	0.715232	0.89769	0.917492
goal4	0.913907	0.950495	0.970297

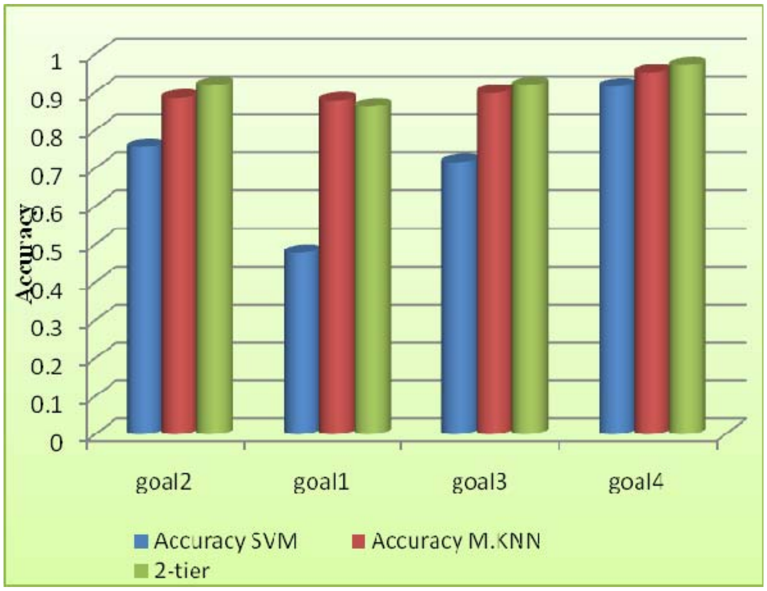

Fig. 2: Accuracy graph between SVM, M.KMM and 2tier proposed Method.

VI. Conclusion

Medical data diagnosis is a dependable application that makes use of data mining techniques to achieve success. When the physicist makes a diagnosis that is characterized by human skill, there is a chance it may fail. On the contrary, data mining may conscript the acquired information from huge amounts of clinical data through data mining and manufacture a predictive model, which can then be used to complete the diagnostic task by employing the classification task. There are a variety of strategies available in this subject for producing the classifier. This dissertation presented a two-tiered system to categorising heart disease in order to improve patient outcomes. Tests using machine learning data sets from the UCI repository are conducted in order to authenticate the proposed methodology. We chose only 14 datasets for the prophecy of a heart disease patient out of a total of 76 datasets available. This projected prediction model assists clinicians in completing the heart disease diagnostic or detection process in a more efficient manner while using fewer features. This disease is most commonly encountered in India, namely in the state of Andhra Pradesh. The sensitivity, specificity, and accuracy of the proposed approach are all considered during the analysis. The simulation result of the accuracy parameter of our projected technique is approximately 95%, which is much higher than the SVM method and somewhat higher than the M.KNN method. When meaningful information can be extracted from these data bases, it can lead to the discovery of rules that can be used to develop later-stage diagnosis tools. Medical data sets are extremely large in size by their very nature. The majority of associative classification algorithms, such as Apriori, use exhaustive search methods to create a large number of rules from which a set of high-quality rules is selected in order to design an efficient classification process.

References:-

[1] K. Manimekalai "Prediction of Heart Diseases using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2016, ISSN(Online): 2320-9801

[2].Lamia AbedNoor Muhammed "Using Data Mining technique to diagnosis heart disease", In proceeding of IEEE,2012.

[3].Ranganatha S., Pooja Raj H.R., Anusha C., Vinay S.K. "Medical Data Mining And Analysis For Heart Disease Dataset Using Classification Techniques", In proceeding of IEEE, 2013.

[4].G.Vaishali, V.Kalaivani "Big Data Analysis for Heart Disease Detection System Using Map Reduce Technique", In proceeding of IEEE, 2016.

[5].Ankita Dewan, Meghna Sharma "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", In proceeding of IEEE 2015.

[6].B.Venkatalakshmi, M.V Shivsankar "Heart Disease Diagnosis Using Predictive Data mining", International Conference on Innovations in Engineering and Technology (ICIET'14) On 21st&22ndMarch, Volume 3, Special Issue 3. In proceeding of IJIRSET.

[7].S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013, no. ICT, pp. 1227–1231.

[8].A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," International Journal of Engineering and Computer Science, vol. 2, no. 9, pp. 1663–1671, 2013.

[9].M. Jabbar, P. Chandra, and B. Deekshatulu, "Cluster Based Association Rule Mining For Heart Attack Prediction," Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.

[10].P. Chandra, M.. Jabbar, and B.. Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628–634.

[11].Mai Shouman, Tim Turner, and Rob Stocker "Using Data Mining Technique in Heart Disease Diagnosis and Treatment", 2012 Japan-Egypt Conference on Electronics, Communications and Computers 978-1-4673-0484-9/12@ 2012 IEEE.pp.173-177

[12].Dhanashree S. Medhekar, Mayur P. Bote,Shruti D. Deshmukh "Heart Disease Prediction System using Naïve Bayes", International Journal of Enhanced Research In Science Technology & Engineering, ISSN NO: 2319-7463 VOL. 2 ISSUE 3, MARCH.-2013.pp.1-5

[13].Vikas Chaurasia and Saurabh Pal "Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739.

[14].Deepali Chandna"Diagnosis of Heart Disease Using Data Mining Algorithm", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, pp.1678-1680.

[15].Aqueel Ahmed, Shaikh Abdul Hannan "Data Mining Techniques to Find Out Heart Diseases: An overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012, pp.18-23.

[16].Mayuri Takore, Prof.R.R. Shelke "Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal for Research in Applied Science & Engineering Technology (IJRASET).

[17].Blake, C.L., Mertz, C.J.: "UCI Machine Learning Dataset", http://mlearn.ics.uci.edu/databases/heartdisease/, 2004.