SURVEY ON SINGLE AND HYBRID MINING APPROACHES IN HEART DISEASE DIAGNOSIS

S.Jayanthi[1], M.Revathi[2]

Assistant Professor, Department of CSE, Agni College of Technology, Chennai.

Abstract: Heart disease is the leading cause of death in US."Heart disease" refers to several types of heart conditions. The most regular type is coronary artery heart disease, which can cause heart attack. Other kinds of heart disease may involve the valves in the heart, or the heart not pump well and cause heart failure. Few people are born with heart disease. Heart disease diagnosis is a important and critical task which can provide prediction about the heart disease so that treatment also made easy. Here, data mining played a vital role in diagnosis of heart disease with improved value. Based on this perception, many researches have been conducted recently. So, analyzing those diagnosis techniques may lead to new improvement in this area. Consequently, we present a detailed survey of standard journals. The survey of the papers related to heart disease .From the survey the finding is that multiple mining techniques contribute more effectiveness than single data mining techniques and some techniques have obtain more than 80% precision. Ultimately, some of the research issue is also addressed to lead the further research on the same path.

Keywords: *Heart disease, Single mining Technique, Hybrid mining technique*

1. INTRODUCTION

Heart diseases are one of the life- threatening diseases and it has more impact on human health. So various research works have undertaken to provide more attention in heart disease diagnosis .Various types of heart diseases was discussed and founded how they direct to heart attack [32]. Health related decision was enabled by knowing about the structure and performance of the heart. New born babies also have the risk of heart disease. The symptoms in people vary depending on the type of heart disease. For many people, chest discomfort or a heart attack is the first sign. It is possible to occur while the heart does not meet the circulatory demands of the body [22]. The physician takes decision based on the

Patient's answers to questions and lab results and experience [15]. When block occurs in coronary arteries blood flow to the heart muscles was decreased. He can perform several tests to diagnose heart disease, including chest X-rays, coronary angiograms, electrocardiograms (ECG or EKG), and exercise stress tests. In advance of medical and surgical action the patient with heart disease reached adulthood [38]. There are various diseases that affect the heart and arteries but MI, renal failure and stroke are more prevalent. Myocardial infarction was caused due to the block in coronary blood vessels which leads to decrease of blood supply in heart muscles. Strokes occurred as a result of blockage in the blood flow to the brain linked to a hemorrhage or a blockage of the arteries that supply blood to the brain. There is a wide range of long-term consequences: heart failure, stroke, kidney failure etc.

The different types of heart disease widely in the world are Coronary heart disease, Heart failure ,Coronary artery disease, Ischemic heart disease, Cardiovascular disease, Atherosclerosis, Chronic obstructive pulmonary disease, Congenital heart disease, Valvular heart disease .Mostly heart attacks are occurred when the plaque on the artery ruptures and a clot then forms, stopping blood flow. And the diagnosis of heart disease was based on both clinical findings and ecg and radiological investigation.. Correct diagnosis of the heart patient was deferred due to many problems. Diagnosis of heart disease was more expensive and finest decision path finder was used in terms of diagnostic accuracy while minimizing cost in diagnosis [17]. Heart disease can strike suddenly and quick and correct decisions have to be made. Prediction of heart diseases can provide some useful information about the health of patient and also made treatment easy. The prediction can be made with many computer aided diagnosis methods. The latent of computer aided tools for medical decision making was realized half a century ago [11], and many algorithms have been developed to build MDS applications for a variety of medical specialties. Some applications using single technique and some other applications using multiple techniques to diagnosis the heart disease. Data mining tools provide successful results in disease diagnosis.

2. SINGLE AND HYBRID DATA MINING TECHNIQUES IN HEART DISEASE DIAGNOSIS

Various data mining techniques have been applied to help health care professionals in the diagnosis of heart disease. Those most frequently used focus on classification: naïve bayes, decision tree, and neural network. Other data mining techniques are also used including kernel density, automatically defined groups, bagging algorithm, boosting algorithm and support vector machine.

Recently, researchers started using hybrid data mining techniques in the diagnosis of heart disease. K.C. Tan and E.J. Teoh *et al,* [4] have proposed a hybrid approach consist of two predictable machine learning algorithms. Genetic Algorithms (GAs) and Support Vector Machines (SVMs) were the two proposed algorithm combined effectively based on a wrapper approach. The proposed method showed84.07% accuracy.

Humar Kahramanli and Novruz Allahverdi, [26] have developed a hybrid neural network which included Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN). The proposed method achieved accuracy values of 84.24% and 86.8%.

Table 2 illustrates a sample of data mining techniques used in the diagnosis of heart disease

Type	**Author**	***Technique***	Accuracy
Single	Tu, et al.	Bagging algorithm	81.4%

Single	Paolo Melillo	Classification and regression tree(CART	85.4%
Hybrid	K.C.Tan and E.J. Teoh	Genetic algorithm and Support Vector Machine	84.07%
Hybrid	Kemal Polat and Salih Gunes	Fuzzy logic and AIRS	92.59

Survey of single and hybrid data mining techniques in the diagnosis of heart disease shows different accuracies, with the hybrid techniques showing better accuracy than single techniques. The best accuracy achieved using single data mining technique is 85.4% by naïve bayes [23]. However, the best accuracy achieved using hybrid data mining technique is 92.59% by AIRS. Hybridized data mining techniques are enhancing the accuracy of heart disease diagnosis.

3. SURVEY OF HYBRID APPROACHES IN HEART DISEASE DIAGNOSIS

In this section, many research articles related to heart disease have been reviewed. K.C. Tan and E.J. Teoh *et al,* [4] have proposed a hybrid approach consist of two predictable machine learning algorithms. Genetic Algorithms (GAs) and Support Vector Machines (SVMs) were the two proposed algorithm combined effectively based on a wrapper approach. Here, by an evolutionary process genetic algorithm component searches for the best feature data set. Based on the feature subset represented by GA, the SVM classified the patterns into reduced data set. This cyclic method was known as wrapper approach. UCI machine learning repository provided 5 set of data and it was checked by the proposed GA and SVM hybrid approach. After that the data was combined with some of the conventional classifier in the data mining community and showed that the collected result of hybrid approach provided a high standard classification.

Also the consistency of the GA-SVM hybrid was clearly seen from the histogram analysis and box plots. The hybrid approach included the utilization of a correlation measure to improve the average fitness of a chromosome population and the substitution of weaker chromosomes based on the correlation measure improved the ability of hybrid classification. The analysis demonstrated GA-SVM hybrid as a good classifier when the irrelevant attributes were removed. The GA-SVM hybrid approach attained an average accuracy of 76.20% which was relatively high. The robustness of the GA-SVM hybrid in the multi-class domain was showed by the obtained average accuracy 84.07%.

Kemal Polat and Salih Gunes, [24] have offered a hybrid approach based on attribute selection, fuzzy Weighted preprocessing and Artificial Immune Recognition System (AIRS) to medical decision support systems.The hybrid approaches based on attribute selection have two steps. The dimensions of heart disease and hepatitis disease datasets were reduced to 9 from 13 and 19 in the attribute selection (AS) sub-program by means of C4. 5 decision tree algorithm. The second step was heart disease and hepatitis disease datasets were normalized in the range of

[0, 1] and were weighted via fuzzy weighted pre-processing. The obtained classification accuracies were 92.59% and 81.82% using 50–50% training-test split for heart disease and hepatitis disease datasets. AIRS have proved an effective performance on many problems such as machine learning benchmark problems and medical classification problems like breast cancer, diabetes and liver disorders classification. They have used the heart disease and hepatitis disease datasets taken from UCI machine learning database as medical dataset.

Humar Kahramanli and Novruz Allahverdi, [26] have developed a hybrid neural network which included Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN). The proposed method accuracy, sensitivity and specificity measures were evaluated which were used commonly in medical classification. The aim of classification was to increase the reliability of the results obtained from the data. Here a new method was Presented for classification of data of a medical database. The proposed algorithm achieved the highest accuracy rate when comparing the records in the UCI web site and related previous studies for diabetes dataset. The proposed method achieved accuracy values of 84.24% and 86.8% for Pima Indians diabetes dataset and Cleveland heart disease dataset respectively. The classification accuracies obtained by the proposed hybrid neural network were one of the best results compared with the results reported in the literature.

Mai Shouman, Tim Turner, Rob Stocker have proposed a model to systematically close those gaps to discover if applying hybrid data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.In this paper data mining classification techniques were used to diagnosis the heart disease and its treatment.

4. SURVEY OF HYBRID APPROACHES IN HEART DISEASE DIAGNOSIS

Carlos Ordonez, [1] introduced an new algorithm to minimize the number of rules . The introduced algorithm searches for association rules in a training set and finally validates them on an independent test set. In medical terms, to the degree of disease in four specific arteries, the association rules related heart perfusion measurements and risk factors. Association rules were applied on a data set containing medical records of patients with heart disease. Search limitations and test set validation importantly minimized the number of association rules and formed a set of rules with high predictive accuracy.

Unfortunately, when association rules were applied on a medical data set, they produced an extremely large number of rules. They used the train and test approach which used two disjoint samples from a data set to search and validate rules. To filter rules on the test set, support, confidence and lift have different importance. They opinioned that to validate rules confidence was the most important metric. Based on heart perfusion measurements and risk factors they used association rules to predict the degree of narrowing in four arteries.

They presented medically significant rules discovered on medical data set that remain valid in several independent train/test cycles. The two problems were addressed such as large numbers of rules were obtained by the standard association rule algorithm and the validation of rules on an independent set, which was required to eliminate unreliable rules.

Kemal Polat and Salih Gunes, [7] have introduced a feature selection method called Kernel F-score Feature Selection (KFFS) which is used as pre-processing step in the classification of medical datasets. The proposed KFFS method has two stages. In first stage by means of Linear (Lin) or Radial Basis Function (RBF) kernel functions, the features of medical

datasets have been transformed to kernel space. Using F-score formula, the F-score values of medical datasets with high dimensional feature space have been calculated. The cause of using kernel functions transformed from non-linearly separable medical dataset to a linearly separable feature space. To evaluate the performance of KFFS method the UCI (University California, Irvine) machine learning database used were heart disease dataset, SPECT (Single Photon Emission Computed Tomography) images dataset and Escherichia coli Promoter Gene Sequence dataset. The area under ROC curve values (AUC) values obtained from just Least Square Support Vector Machine (LS-SVM) and Artificial Neural Network (LANN)
classifiers without KFFS method on the classification of heart disease .They compared and found the best expert system based on the classification used in medical data set.

Yoon-Joo Park and Se-Hak Chun, *et al,* [18] have proposed a Cost-Sensitive Case-Based Reasoning (CSCBR), a new knowledge extraction technique. It included unequal misclassification cost into conventional case based reasoning. To classify the absence and presence of disease genetic algorithm was used. An effort was taken to minimize misclassification error costs into CBR by the best classification of boundary point and Number of neighbor. A fixed number of nearest neighbors in CBR was overcome by CSCBR. The absence and presence of disease was classified by adjusting the optimal cut-off classification point and cut-off distance point for selecting best neighbors. The CSCBR technique was applied in five medical data sets and then compared the result with C5.0 and CART. The total misclassification cost of CSCBR was lower than other cost-sensitive methods and was originally designed to classify binary case.

Laercio Brito Gonçalves and Marley Maria Bernardes Rebuzzi Vellasco*, et al*, [23] have determined that the Inverted Hierarchical Neuro-Fuzzy Binary Space Partitioning (HNFB-1) was based on the Hierarchical Neuro-Fuzzy Binary Space Partitioning Model (HNFB) which gave an idea that recursive partitioning of the input space. It was able to generate its own structure automatically and allowed a greater number of inputs. The Classification task of HNFB-1 has been evaluated with different benchmark databases such as heart disease data sets. They introduced an Inverted Hierarchical Neuro-Fuzzy BSP System. It was a neuro-fuzzy model which has been specifically created for record classification and rule extraction in databases.

It allowed the extraction of knowledge in the form of interpretable fuzzy rules. Fuzzy accuracy and Fuzzy coverage were the two fuzzy
evaluation measures defined for the process of rule extraction in the HNFB-1 model. The HNFB-1 model had showed better classification performance when compared with several other pattern classification models and algorithms and the processing time converged by HNFB-1 was very less.

Paolo Melillo, Nicola De Luca, Marcello Bracale, and Leandro Pecchia,et al have developed an automatic classifier for risk assessment in patients suffering from congestive heart failure (CHF). The proposed classifier separates lower risk patients from higher risk ones, using standard long-term heart rate variability (HRV) measures. Patients are labeled as lower or higher risk according to the New York Heart Association classification (NYHA). A retrospective analysis on two public Holter databases was performed,analyzing the data of 12 patients suffering from mild CHF (NYHA I and II), labeled as lower risk, and 32 suffering from severe CHF (NYHA III and IV), labeled as higher risk. Only patients with a fraction of total heartbeats intervals (RR) classified as normal-to-normal (NN) intervals (NN/RR) higher than 80% were selected as eligible in order to have a satisfactory signal quality. Classification and regression tree (CART) was employed to develop the classifiers. A total of 30 higher risk and 11 lower risk patients were included in the analysis. The proposed classification trees achieved sensitivity and a specificity rate of 93.3% and 63.6%, respectively, in identifying higher risk patients.

5. FUTURE RESEARCH DIRECTIONS

After analyzing the survey, the following are some of the issues identified that can be taken further to do the research. Even though many techniques have been used in the literature, still there is a need of good techniques to solve the following research issues.

- Data cleaning is an important problem for heart disease database . So handling of missing values for diagnosis problem is a challenging task.
- The heart database has multidimensionality, so identification of significant attributes for better diagnosis of heart disease is very challenging task.
- Most of classification algorithms are not suitable to handle large dataset for diagnosis.
- Selection of most suitable sample of data for classification is another risk for getting better diagnosis.
- Selecting the suitable classification techniques with less computation complexity without affecting its effectiveness is another issue.
- The heart disease database is very sensitive so that much consideration and attention is need in accuracy of diagnosis.

6.CONCLUSION

A survey of many papers related to heart disease diagnosis published in the standard Journals of IEEE, Elsevier, springer and inderscience are presented. From the review, the identification was that hybrid techniques contribute more effectiveness. Also from the accuracy perspectives, hybrid techniques provide more than 90% accuracy as compared with the other techniques presented in the literature. Finally, some of the research issue is also addressed to precede the further research in the same path.

References

[1] Carlos Ordonez," Association Rule Discover with the Train and Test Approach for the Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, PP. 334-343, April 2006

[2] Mu-Jung Huang and Mu-Yen Chen *et al,* "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis" Journal of Expert Systems with Applications, Vol. 32, PP.856–867, 2007

[3] K.C. Tan and E.J. Teoh *et al,* "A hybrid evolutionary algorithm for attribute selection in data mining", Journal of Expert system with applications, Vol.36, PP.8616-8630, 2009

[4] Jesmin Nahar and Tasadduq Imam *et al,*" Association rule mining to detect factors which contribute to heart disease in males and females", Journal of Expert Systems with Applications Vol.40, PP.1086–1093, 2013

[5] Kemal Polat and Salih Gunes," A new feature selection method on classification of medical datasets: Kernel F-score feature selection",Journal of Expert Systems with Applications, Vol. 36, PP.10367–10373, 2009

[6] Resul Das and Ibrahim Turkoglu, *et al,* "Effective diagnosis of heart disease through neural networks ensembles", Journal of expert system with applications, Vol.36, PP. 7675–7680, 2009

[7] K.Rajeswari and V.Vaithiyanathan, *et al*, "Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks ", International Symposium on Robotics and Intelligent Sensors, Vol.41, PP. 1818–1823, 2012.

[8] Akin Ozcift and Arif Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms", Journal of Computer Methods and Programs in Biomedicine, Vol.104, PP.443-451, 2011.

[9] Chih-Lin Chi and W. Nick Street, *et al,* "A decision support system for cost-effective diagnosis", Journal of Artificial Intelligence in Medicine, Vol.50, PP. 149-161, 2010.

[10] Yoon-Joo Park and Se-Hak Chun, *et al,* "Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis", Journal of Artificial Intelligence in Medicine, Vol.51, PP.133-145, 2011

[11] Debabrata Pal and K.M. Mandana, *et al,* "Fuzzy expert system approach for coronary artery disease screening using clinical parameters", journal of knowledge based system, Vol.36, PP.162-174, 2012

[12] Ismail Babaoglu and Og˘uz Findik *et al,* "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine", Journal of Expert System With Applications,
Vol.37, PP.3177-3183, 2010

[13] Jesmin Nahar and Tasadduq Imam, *et al,* "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach", Journal of Expert System with Application, Vol.40, PP.96-104, 2013

[14] Kemal Polat and Salih Gu nes, "A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS", Journal of Computer Methods and Programs in Biomedicine, Vol.88, PP.164-174, 2007

[15] Kemal Polat and Seral S ahan *et al,* "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbor) based weighting preprocessing", Journal of expert system with
applications, Vol.32, PP.625-631, 2007

[16] Humar Kahramanli and Novruz Allahverdi, "Design of a hybrid system for the diabetes and heart diseases", Journal of Expert Systems with Applications, Vol. 35, PP. 82–89, 2008

[17] Nazri Mohd Nawi and Rozaida Ghazali *et al,* "The Development of Improved Back-Propagation Neural Networks Algorithm for Predicting Patients with Heart Disease", In proceedings of the first international conference ICICA, Vol.6377, PP.317-324,2010

[18] Li-Na Pu and Ze Zhao, *et al,* "Investigation on Cardiovascular Risk Prediction Using Genetic Information", Journal of IEEE
Transactions On Information Technology In Biomedicine, Vol. 16, No. 5, September 2012

[19] R. Pfister and D. Barnes *et al,* "Individual and cumulative effect of type 2 diabetes genetic susceptibility variants on risk of coronary
heart disease", Journal of Diabetologia, Vol.54, PP.2283-2287, 2011

[20] R. Goekmen Turan and I. Bozdag *et al,* "Improved Functional Activity of Bone Marrow Derived Circulating Progenitor Cells After

Intra Coronary Freshly Isolated Bone Marrow Cells Transplantation in Patients with Ischemic Heart Disease", Journal of stem cell
review and report, Vol.7, PP.646-656, 2011

[21] Petra A. Karsdorp and Merel Kindt *et al,* "False Heart Rate Feedback and the Perception of Heart Symptoms in Patients with Congenital Heart Disease and Anxiety", International Journal of behavioral Medicine, Vol.16, PP.81-88, 2009

[22] Giorgio Barbareschi and Robbert Sanderman *et al,* "Socioeconomic Status and the Course of Quality of Life in Older Patients with Coronary Heart Disease", International Journal of behavioral Medicine, Vol.16, PP.197-204, 2009
Gayathri. P et.al / International Journal of Engineering and Technology (IJET)
ISSN : 0975-4024 Vol 5 No 3 Jun-Jul 2013 2957

[23] Keyue Ding and Kent R Bailey *et al,* " Genotype-informed estimation of risk of coronary heart disease based on genome-wide association data linked to the electronic medical record", International journal of BMC cardiovascular Disorders, Vol.11,2011

[24] Daisy JA Janssen and Emiel FM Wouters *et al,* "Self-perceived symptoms and care needs of patients with severe to very severe chronic obstructive pulmonary disease, congestive heart failure or chronic renal failure and its consequences for their closest relatives: the research protocol", Journal of BMC palliative care, Vol.7, 2008

[25] Shou-En Lu and Gloria L Beckles *et al*, "Evaluation of risk equations for prediction of short-term coronary heart disease events in patients with long-standing type 2 diabetes: the Translating Research into Action for Diabetes", International Journal of BMC Endocrine Disorders, Vol.12, 2012.

[26] Lucile Houyel and Babak Khoshnood *et al,* "Population-based evaluation of a suggested anatomic and clinical classification of congenital heart defects based on the International Paediatric and Congenital Cardiac Code", International Journal of rare diseases, Vol.6, 2011.

[27] Evanthia E. Tripoliti and Dimitrios I. Fotiadis *et al,* "Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm", Journal of IEEE Transactions On Information Technology In Biomedicine, Vol. 16, No. 4, July 2012

[28] V. Sree Hari Rao and M. Naresh Kumar, "Novel Approaches for Predicting Risk Factors of Atherosclerosis" Journal of IEEE Journal Of Biomedical And Health Informatics, Vol. 17, No. 1, January 2013.

[29] Zhihua Cui *et al,* "Training artificial neural networks using APPM", International Journal of wireless and mobile computing, Vol.5, PP.168-174, 2012

[30] P.K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules and decision tree rules", Journal of Computer Sciences, Vol.24, PP. 27–40, 2012.