International Classification and Retrieval of Heart Disease from Unstructured Big Medical Data of Health Care System using Deep Neural Network

Uma Jothi T., M.Sc (CS)., M.Phil (CS)., Madurai, India, adlinjothi@gmail.com

Abstract - **This research is aimed at international classification and retrieval of the heart disease from massive unstructured medical data from Health Care System by applying Deep Neural Network methodologies. The International classification of diseases (ICD) is the most recognized diagnostics tool for medical classification being used by Healthcare professionals. The basic concept of ICD is founded on the standardization of the nomenclature for the names of diseases and their basic systematization in the hierarchically structured and unstructured category. There are lot of advantage that this classification offer but it is unlikely to find appropriate code for some disease in order to provide adequate diagnosis due to shortage of sufficient data especially to a patient that comes first time to the ambulance or hospital. Soto bring an efficient and accurate classification, Deep Neural Network approach is used in this research that can be applied for all circumstances. The proposed framework evaluates a query in four phases. In phase 1, structured data is used to filter the clinical data warehouse. In phase 2, feature extraction modules are executed on the unstructured medical data in a distributed manner to complete the query. In phase 3, the deep neural network algorithm is applied on to the extracted features for recognize the heart disease pattern. In phase 4, the most relevant heart disease of the query is retrieved from the big medical data.**

Keywords — Heart Disease Classification, Feature selection, Feature extraction, Deep Neural Network, GenPCNN and CNN

I. Introduction

Heart is one of the most significant organs in human's body that plays a vital role in supplying oxygen and nutrients in blood flow to itself and to other organs for their survival through blood vessels. Heart disease describes a number of conditions that affect the heart and are the leading cause of death worldwide. Together heart disease and stroke are the most widespread health problems the countries face for long days. Identifying the heart disease is a challenging task that requires several biological indicators and risk factors including age, gender, high blood pressure, diabetes, cholesterol level and many other clinical indicators.

In the healthcare sector, Data mining techniques have provided an effective solution for predicting the different type diseases to identify some diagnostic traits that are essential for the patients. Diagnosis is a complex task requiring a lot of skill and experience and, in many cases, the diagnosis is based on current patient and physician experience. Health data mining is a challenge that includes many misconceptions and uncertainties; with the help of mining techniques, the number of tests should be condensed and this reduction on diagnosis plays a key role in timing and accuracy. This paper aims to illustrate data mining techniques for heart disease prognosis.

An increasing number of patients with heart disease worldwide have urged researchers to conduct extensive research to show hidden patterns in clinical data. This section provides an overview of previous sample studies for heart patterns. Not only different techniques, but different data sets on heart disease are considered to fair comparisons. Finally, the gap in existing literature as the main motivation of the study was also demonstrated. Some important studies are:

Das et al. Neurobiologists have been advised to diagnose heart disease. A basic approach combines a new pattern with a combination of probabilities or predictive values from many creators. Based on the developed experimental model, the accuracy of the 97.4% test assay was found on a sample of 215 samples [1]. Pandey et al. suggested rendering of group algorithms using data set for heart disease. They evaluate the performance and accuracy of some algorithms for grouping. Cluster performance is calculated using clustered clusters. Finally, they suggested that the author of the right-of-the-range, 85% predictive, positive-density cluster was the most universal algorithm for heart disease diagnosis [2].

Karaolis et al. established a data collection system using a joint analysis based on Apriori algorithm to evaluate cardiac risk factors with WEKA devices. A total of 369 cases from the IBS poll in Paphos were collected, with most of them having more than one event. The laws chosen are determined according to the importance of each law. Each quote rule is further evaluated by examining the number of cases in the database [3]. Nidhi Bhatla, Kiran Jyoti and others. [4]. , Proposing a logical fate with the trees of decision and naivete Bayes. In this observation, six traits are reduced to four attributes, its aid in reducing the tests that

DOI : 10.35291/2454-9150.2020.0547

are taken by the patient. Indira S.FalDessai et al. [5], the technique proposed for the intelligent heart rate prediction system. In this research, the network of neurons may be used. This provides a higher accuracy of 92.10% compared to other distribution methods. Mr.Akhil Jabbar, Dr.BLDeekshatulu and Dr. Priti Chandra and others. [6]. An applied analysis of the main components in Lazy's classification to form a group association law would be used to predict the onset of heart disease. He gives 90% of accuracy and affirms that men are more influential than women.

Purushottam Professor Kanak Saxena and Richa Sharma and others. [7]. , KEEL is a Java open source tool for evaluating evolutionary algorithms for data collection problems. It is expected that this system has great potential for predicting heart disease risk factors such as S. Sugana and P. Tamie Selvey and others. [8]. They are J48, Bayes Net, Naive Bayes, and conventional sabotage algorithms and offer REPTREE algorithms to reduce the uncertainty of death and predict the accuracy of the results and to produce the results corresponding to the heart disease predictions. Durairaj.M and Revathi.V et al. [9] In this paper, a multipolar propagation algorithm is proposed, which is implemented in matlab 7.0. The MLPA algorithm is best when compared to other algorithms and uses the TRAINBR function 96.30% of the highest accuracy compared to all of the training functions of the advanced reaction algorithm. This article suggests that MLPA can be an effective tool for predicting heart disease with high accuracy.

Aswathy Wilson, Gloria Wilson, Likhiya Joy .K introduced the WAC and K Means weighting materials team [10]. These techniques can be used with the standard data standardization process (CRISP-DM), which is used to improve efficiency, classification and efficiency of data. After the K-Means rating is more accurate than the weight divider. Ankur Makwana and Jaymin Patel et. [11] Requested an integrated combination of Bayes and Gene algorithms. Following the law, the analysts are compared to the traditional tree-oriented algorithmic algorithms based on precision and accuracy. In this study, the results are more accurate than other data mining techniques.

According to Ordonez [15] the heart disease can be predicted with some basic attributes taken from the patient and in their work have introduced a system that includes characteristics of an individual human being based on totally 13 basic attributes like sex, blood pressure, cholesterol and others to predict the likelihood of a patient getting affected by heart disease. They have added two more attributes i.e. fat and smoking behaviour and extended the research dataset. The data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network are utilized to make predictions and the results are analysed on Heart disease database. Yılmaz, [25] have proposed a method that uses least squares support vector machine (LS-SVM) utilizing a binary decision tree for classification of cardiotocogram to find out the patient condition.

Duff, et al. [16] have done a research work involving five hundred and thirty-three patients who had suffered from cardiac arrest and they were integrated in the analysis of heart disease probabilities. They performed classical statistical analysis and data mining analysis using mostly Bayesian networks. Frawley, et al. [17] have performed a work on prediction of survival of Coronary heart disease (CHD) which is a challenging research problem for medical society. They also used 10-fold cross-validation methods to determine the impartial estimate of the three prediction models for performance comparison purposes. Lee, et al.[18] proposed a novel methodology to expand and study the multi-parametric feature along with linear and nonlinear features of Heart Rate Variability diagnosing cardiovascular disease. They have carried out various experiments on linear and non-linear features to estimate several classifiers, e.g., Bayesian classifiers, CMAR, C4.5 and SVM. Based on their experiments, SVM outperformed the other classifiers.

Noh, et al.[19] suggested a classification method which is an associative classifier that is constructed based on the efficient FP growth method. Because the volume of patterns can be diverse and huge, they offered a rule to measure the cohesion and in turn allow a tough choice of pruning patterns in the pattern-generating process. Parthiban, et al. [20] have proposed a new work in which the heart disease is identified and predicted using the proposed Coactive Neuro-Fuzzy Inference System (CANFIS). Their model works based on the collective nature of neural network adaptive capabilities and based on the genetic algorithm along with fuzzy logic in order to diagnose the occurrence of the disease. The performance of the proposed CANFIS model was evaluated in terms of training performances and classification accuracies. Finally, the results show that the proposed CANFIS model has great prospective in predicting the heart disease.

Guru, et al. [21] have proposed the computational model based on a multilayer perceptron with three layers is employed to enlarge a decision support system for the finding of five major heart diseases. The proposed decision support system is trained using a back propagation algorithm amplified with the momentum term, the adaptive learning rate and the forgetting mechanics.

In recent years, neural network models have presented their outstanding performance for data prediction and tackling various classification problems. Deep learning techniques have played a significant role in the healthcare domain for disease classification like heart disease

The main objective of this paper is

DOI : 10.35291/2454-9150.2020.0547

- To develop a deep learning methods of heart disease classification and retrieval algorithms using machine learning approach.
- To improve the classification accuracy
- To reduce the error rate.

The rest of the paper is organized as follows: Section II discusses the overview of proposed methodology. In Section III, the performance of the proposed method is compared with the existing approaches. Finally, Section IV concludes with a discussion about the proposed methodology.

II. PROPOSED METHODOLOGY

The architecture of the proposed system is illustrated in Fig.1.The major components of this system are Dataset Collection, Preprocessing, Attribute Selection, Feature Extraction and Prediction.

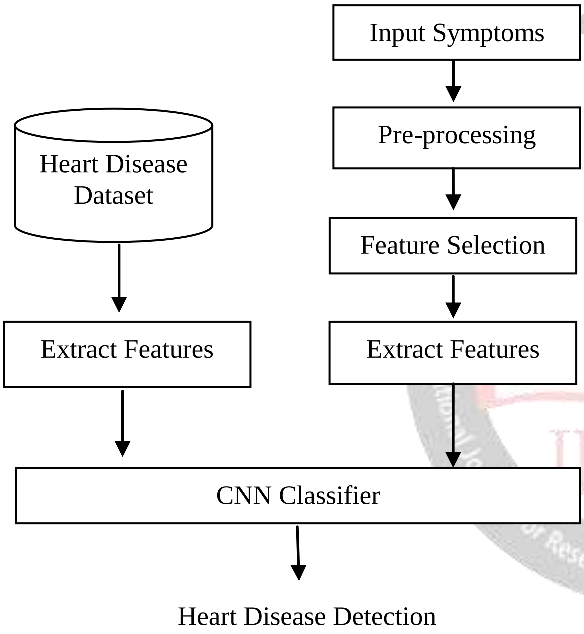

Fig.1. Overall Block Diagram of proposed Method

A. Preprocessing

Preprocessing is an important step in the knowledge discovery process, as real world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation, and data reduction. In this module cleaning and filtering of the data set is done to remove duplicate records, normalize the values, accounting for missing data and removing irrelevant data items. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. In this proposed work, the most probable value is used to fill in the missing values.

B. Attribute Selection using GenPCNN

Automated medical diagnostic approaches are mainly based on a Machine Learning (ML) algorithm. A large number of attributes that can surpass the number of data themselves often characterizes the data used in ML. This problem known as "the curse of dimensionality". It creates a challenge for various ML applications for decision support. This can increase the risk of taking into account correlated or redundant attributes which can lead to lower classification accuracy. Therefore, the process of eliminating irrelevant attribute is a vital phase for designing decision support systems with high accuracy. This thesis proposed a new hybrid feature selection method by combining Genetic and Pulse Coupled Neural Network(PCNN) approaches. In this module the best attribute is selected by using the GenPCNN method. The algorithm of GenPCNN is follows

Input: Attributes (X) and its values, Cross over Probability (CP), Mutation Probability (MP)

Output: Best attribute (S) & Fitness Value (S).

1. Generate random population from each Attribute and Consider as the Initial Population P.
2. Evaluate the fitness f(X) of each attribute (X) in the population (P) by using the below formula.

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where x_i is ith attribute values in the attribute X and N is the total values in the attribute

3. Create a new population (NP) by repeating following steps until the new population (NP) is complete.
4. Select attributes from a population P according to their fitness. If the fitness value is high, then those attributes as selected as best fitness value (W), otherwise rejected.
5. With a crossover probability (CP) cross over the selected attributes to form a new attribute. If no crossover was performed, new attribute is an exact copy of selected attributes.
6. With a mutation probability (MP) mutate selected attribute at each locus.
7. Place selected attribute in a new population (NP).
8. Use new generated population (NP) for a further run of algorithm.
9. If the end condition is satisfied, stop, and return the best solution(S) in current attribute.
10. Go to step 2.
11. Finally the output of the genetic algorithm is the fitness value (W) and the best attribute (S).
12. Give the best attribute (S) as the weight value of the PCNN and execute PCNN.
13. From the PCNN and return the best attribute (S) as the output.

Finally the output of the genetic algorithm is the fitness value and the best population. It is given into the PCNN as the input.

DOI : 10.35291/2454-9150.2020.0547

C. *Feature Extraction*

In this module, the features are extracted from the best attributes. The first order feature such as mean and standard deviation are used. The mean feature is calculated by using the below formula.

$$\mathrm{M} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i)} \qquad (1)$$

where xi is the value of each best attribute and N is the total no of patients. The standard deviation formula is calculated by using the below formula

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (2)$$

Where xi is the value of each best attribute, µ is the mean value and N is the total no of patients. These features are combined to form a single feature vector and they are trained by CNN.

D. *Heart Disease Prediction using Convolution Neural Network*

The proposed method applied a Convolution Neural Network (CNN) as the classifier for heart disease. In this stage the extracted features are given into the CNN for predicting the condition of the Heart.

A CNN consists of one or more convolution layers and pooling layers. The convolution layers play the role of additional feature extractor by applying convolution filter. The pooling layer is to perform down sampling, that is, to reduce the amount of computation time by reducing the extracted features in convolution layer. There are two kinds of pooling layers: max pooling and average pooling. In max pooling, this paper take only the value of the largest pixel among all the pixels in the receptive field of the filter. In the case of average pooling, we take the average of all the values in the receptive field. The output of the pooling layer is given fully connected layer for detecting the heart disease based on the training features.

III. PERFORMANCE ANALYSIS

A. *Data Set Used*

In this module the dataset is collected and stored for processing which was extracted from Health Care Centre for machine learning and intelligent systems. It includes 270 records of patients that indicate several biological indicators including age, gender etc, as described in Table 1 that are used for the proposed work.

Table 1: Attributes of Disease Dataset

S.No	Attribute Name
1.	S.NO
2.	NAME
3.	AGE
4.	GENDER
5.	ADDRESS
6.	CONTACT NO
7.	GTT
8.	FASTING
9.	POST PRA
10.	HBA1C
11.	HEIGHT
12.	WEIGHT
13.	BMI
14.	URIC ACID
15.	TG
16.	LDL-C
17.	VLDL
18.	HDL-C
19.	SYSTOLIC
20.	DIASTOLIC
21.	HSCRP
22.	SMOKERS
23.	CHAIN SMOKERS
24.	NON-SMOKERS
25.	ALCOHOL CONSUMPTION
26.	DIET
27.	PHYSICAL ACTIVITY
28.	WORK STRESS

B. *Performance Analysis*

Number To evaluate the proposed method performance, several performance metrics are available. This paper uses Precision Rate, Recall Rate, Sensitivity, Specificity and F-Measure to analyze the performance.

1. Precision Rate

The precision is the fraction of retrieved instances that are relevant to the find.

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

where, TP = True Positive (Equivalent with Hits)

FP = False Positive (Equivalent with False Alarm)

2. Recall Rate

The recall is the fraction of relevant instances that are retrieved according to the query.

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

where, TP = True Positive (Equivalent with Hits)

FP = False Negative (Equivalent with Miss)

3. F-Measure

F-measure is the ratio of product of precision and recall to the sum of precision and recall. The f-measure can be calculated as,

$$F_m = (1+\alpha) * \frac{Precision*Recall}{\alpha*(Precision*Recall)} \quad (5)$$

4. Sensitivity

Sensitivity also called the true positive rate or the recall rate in some field's measures the proportion of actual positives.

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} \quad (6)$$

where, TP – True Positive (equivalent with hit)

FN – False Negative (equivalent with miss)

5. Specificity

Specificity measures the proportion of negatives which are correctly identified such as the percentage.

$$\text{Specificity} = \frac{TN}{(FP+TN)} \quad (7)$$

where, TN – True Negative (equivalent with correct rejection)

FP – False Positive (equivalent with false alarm)

To analysis the performance of the proposed system, it is compared with various techniques by using the performance metrics which are mentioned above. This is shown in the below graphs.

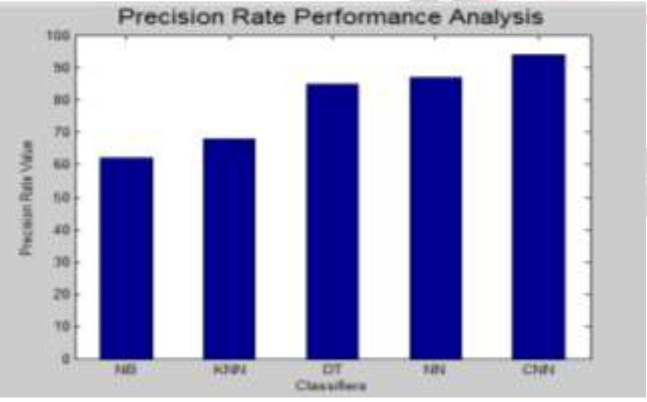

Fig.2. Precision Rate Analysis

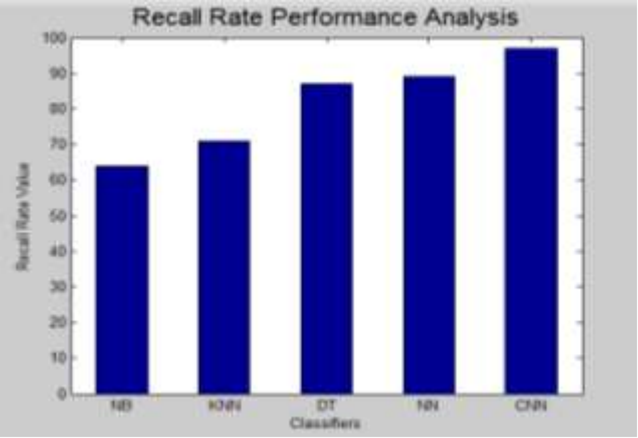

Fig.3. Recall Rate Analysis

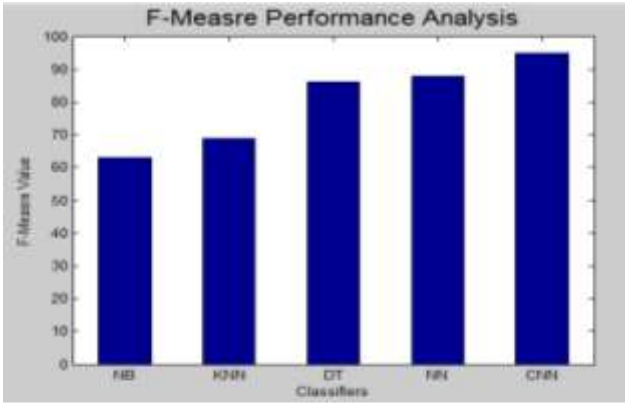

Fig.4. F-Measure Analysis

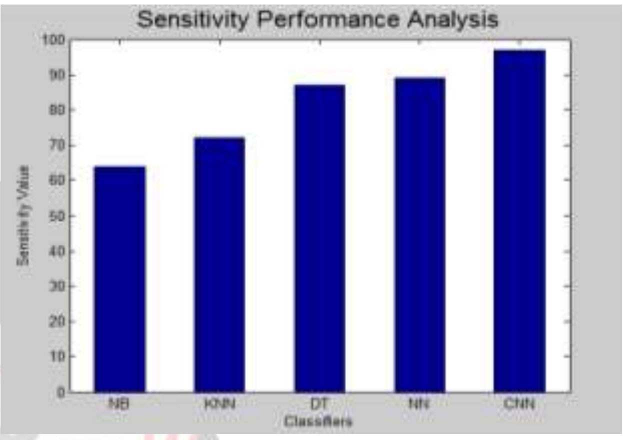

Fig.5. Sensitivity Analysis

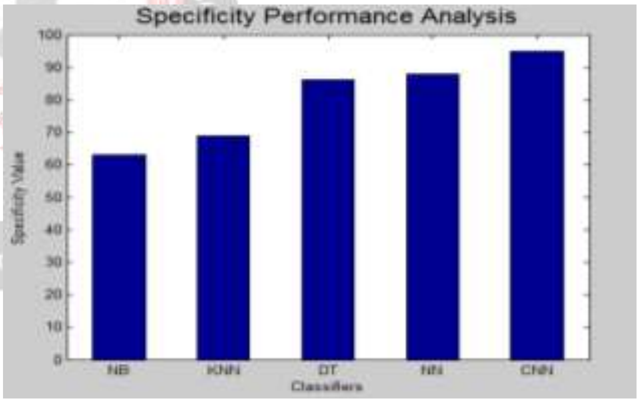

Fig.6. Specificity Analysis

IV. CONCLUSION

In this paper, we discussed various machine learning algorithms which are applied for the prediction and classification of heart disease; each algorithm has some advantages and disadvantages. The comparison was conducted on the basis of the above performance criteria. In this work, we have imposed a decision for prioritizing algorithms that will be useful for medical data. The experimental results reveal that it's found CNN algorithm is doing better than other machine learning algorithms.

REFERENCES

[1] R. Das, I. Turkoglu, and A. Sengur, "Diagnosis of valvular heart disease through neural networks ensembles," Elsevier, 2009.

[2] A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method," International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2277798, Vol 2, Issue10, October 2013.

[3] M. Karaolis, J. A. Moutiris, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events," Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. https://doi.org/10.1109/iembs.2009.5334656.

[4] Nidhi Bhatla and Kiran joyit " A Novel Apporach for Heart Disease Diagonisis using Data mining and Fuzzy logic" International Journal of Computer Applciation,(0975 – 8887),Vol.54.No.17 Sep 2012.

[5] Indira S. FalDessai "Intelligent Heart Disease Prediction System using Probabilistic Neural Netwrok", Internationl Journal on Advance Computer Theory and Engineering,ISSN (2319-2526),Vol:2 Issue:3 2013.

[6] M.Akhil Jabbar, Dr.B.L.Deekshatulu and Dr.Priti Chandra "Heart Disease Prediction using Lazy Associative Classification",IEEE, 2013.

[7] Purushottam, Prof (Dr) Kanak saxena and Richa Sharma "Efficient Heart Disease Prediction system Using Decision Tree",ICCCA 2015(International Conference Computing, Communication and Automation,IEEE,2015.

[8] S.Suganya,P.TamijeSelvy "A Proficient Heart Diseases Prediction method using Fuzzy-CART Algorithm", International Journal of Scientific Engineering and Applied Science(IJSEAS),Vol2 Issue:1 January 2016.

[9] Durairaj.M and Revathi.V "Prediction of Heart Disease Using Back Propagation MLP Algorithm", International Journal of Scientific & Technology Research Volume 4, Issue 08, August 2015.

[10] Aswathy Wilson, Gloria Wilson, LikhiyaJoy.K "Heart Disease Prediction Using the Data Mining Techniques", International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 1, Jan-Feb 2014.

[11] Ankur Makwana and Jaymin Patel "Decision Support System for Heart Disease Prediction using Data Mining Classification Techniques", International Journal of Computer Applications (0975 8887), Volume 117 - No. 22, May 2015.

[12] www.mayoclinic.org/diseases-onditions/heartdisease/basics/definition/con-20034056.

[13] www.allen-temple.org/ministries/education/health-educationministry/2159-2016-february-is-heart-disease-month.

[14] https://the-modeling-agency.com/how-data-mining-is-helping-healthcare.

[15] Carlos Ordonez, "Improving Heart Disease Prediction using Constrained Association Rules", Technical Seminar Presentation, University of Tokyo, 2004.

[16] 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.

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

[18] HeonGyu Lee, Ki Yong Noh and Keun Ho Ryu, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, 2007.

[19] 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.

[20] 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.

[21] Niti Guru, Anil Dahiya and Navin Rajpal, "Decision Support System for Heart Disease Diagnosis using Neural Network", Delhi Business Review, Vol. 8, No. 1, pp. 1-6, 2007.

[22] SellappanPalaniappan 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.

[23] 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.

[24] X. Yanwei et al., "Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease", Proceedings of International Conference on Convergence Information Technology, pp. 868-872, 2007.

[25] Ersen Yilmaz and CaglarKilikcier, "Determination of Patient State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree", Master Thesis, Department of Electrical Electronic Engineering, Uludag University, 2013.

DOI : 10.35291/2454-9150.2020.0547