P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

HYBRID TECHNIQUE FOR ENHANCING THE ACCURACY OF EARLY PREDICTION OF CARDIOVASCULAR DISEASE

Dineshkumar M[1] , Sivakumar D[2] Jeyabalan S[3]
[1]Department of Information Science and Engineering
[2,3] Department of Computer Science and Engineering
Rajarajeswari College of Engineering, Bangalore, India-560074
dskumarcse@yahoo.co.in,
dineshkumar@rrce.org,kavijeyabalan1968@gmail.com

Abstract: Most people actually have heart disease because of work and self depression. Consequently, every year deaths from heart disease are steadily increasing. While new technologies have been developed worldwide, not all of them are successful in diagnosing heart disease well in advance. This paper looked at the early-stage decision-making strategy for heart disease using Decision Tree, k-means, SVM and neural network. This analysis aims primarily to improve the accuracy with hybrid machine learning using the UCI online data repository. Through the results obtained through this research study it is identified that the hybrid machine learning technique DTSVM is provides more accurate result when compared to all other methods. The 88.3 % accuracy is achieved through this hybrid technique and error rate 11.1%.

Keywords: Decision Support System, K-Means, Classification, Decision Tree, SVM, Neural Network.

1. Introduction

Quality treatments at a reasonable rate are the main obstacles for the healthcare industry. A quality service provides the correct diagnosis and effective treatment for the patient. Bad clinical decisions can lead to tragic, inappropriate outcomes. All hospitals have used certain software to predict the condition through techniques for data mining. cardiovascular or cardiovascular disease is the condition caused by a problem with heart and other veins, blood vessels, arteries and veins [1]. Many countries have crisis with an growing number of cardiovascular disorders and this is one of the key causes for men or women's weakening and death in the universe over the age of 65. Cardiovascular disease is widely viewed in many countries as a second epidemic, replacing infectious disease deaths [2]. Indian Medical Research Council (ICMR) figures indicate that almost 25 % of deaths are caused by heart disease in ages 25 to 69 years. In 2008 in particular, almost 50 percent of world wide cardiovascular disease deaths than wounds and other diseases were caused [3]. If it goes on, the death level in 2030, due to cardiovascular disease (CVD) may rise to 76%. The highest

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

group of non-communicable diseases is composed of CVDs. CVD is the only possible explanation for the destruction of the planet. [4].

The root causes of cardio vascular disease are smoking, fitness weaknesses, food insulation, alcohol consumption, excessive stress, high blood pressure, unregulated smoking, poor cholesterol levels, obesity or blood sugar etc. The symptoms and actions of a person who is likely to develop a disease are a danger [5]. However, it can increase the risk of a current illness. There are two types of risk factors: firstly, uncontrollable and secondly controllable risk factors.

The following characteristics are age, sex and family background as uncontrollable factors of risk. The risk factors controllable are smoking, weight, cholesterol, diabetics and blood pressure [6]. In later menopause women and men over 40 years old typically cardiovascular disease; and the majority of cardiovascular disease is recorded for women over 65 years old. Men and women are more likely to have cardiovascular diseases in early years and are more vulnerable if their family members have cardiovascular disease [7].

Blood clots formed in the venous membranes or helped by cigarette smoke chemistry and plaque in the walls of the artery may form. A greater risk of cardiovascular disease if body weight increases. For those that have body fat added in the tail area, this is particularly true. The risk of heart disease can be decreased by certain dietary ingredients. In the blood building of the artery walls, the cholesterol may contribute to heart disease, a type of atherosclerosis [8].

The extent of cardiac diseases in people was assessed by a variety of approaches using data mining and neural networks. Several different approaches, such as k-Nearest Neighbor Algorithm (KNN) for the classification of extreme disease cases, Decision Trees (DT), Genetic algorithm (GA) and Naïve Bayes (NB) have been implemented. Heart disorders are very difficult and must therefore be treated carefully. This cannot have an effect on the heart or result in early death. Data processing mining is an important aspect of predicting and interpreting outcomes of heart disease. Use the famous Cleveland data set from a UCI learning repository for experimental validation. The best way to predict heart disease and brain disease are neural networks. There are 13 prediction attributes to the suggested method. The results demonstrate an increased degree of productivity relative to existing methods.

Carotid Artery Stenting (CAS) is also common in recent years for medical care. The CAS leads to major cardiovascular adverse events (MACE) in elderly cardiovascular patients. Their assessment is very important. Neural network models are implemented that incorporate not only later probabilities but also several expected values. In contrast to previous works, the model achieves a precision rating of 89.01%. The neural network (NN) used to improve the efficiency of predictive cardiac illnesses for all experiments in Cleveland.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

2. Literature Survey

Dangare et.al proposed [13] a prediction system for heart disease containing colossal data measurements used to isolate invisible data for in-depth therapeutic detection. The standard goal is to build a cardiac prediction system for the identification of cardiovascular disease using the coronary disease dataset collected. Critical pattern extraction from the information distribution center for coronary heart disease has been anticipated.

N. Deepika and K. Chandrashekar proposed [14] that the information storage facility be prepared at first to make the mining process more powerful and that it be used in the dataset after that affiliation administrator. In the light of the medical guidelines on heart attack details they are linked by equal interim binning with the expected value.

Niti Guru et.al suggested [15] that neural patient prediction networks and research should be carried out. These tests were conducted on a reviewed patient report information collection. For analysis of coronary heart disease, the monitoring system was educated. The planning was completed by the back propagation algorithm..

Asghar, S has proposed [16] that in recent years information mining procedures have become one of the actual fields of inquiry for the removal of unique and useful data. People are able to effectively appreciate these data. This data extraction has been physically registered and evaluated using measurable systems.

Parisa Naraei et.al investigated [17] the accuracy of neural perceptron systems in multi layer and vector supporter for cardiovascular data collection. A dataset of 303 patients has also dissected the adequacy of supporting vector machinery. The analysis shows that multilayer perceptive neural networks could be much more reliably defined by Bolster vector machinery.

Deepali Chandna has suggested [18] that the amount of research required can be reduced data mining techniques. Cardiovascular disease is the leading cause for decreased disease and in an unpleasant operation coronary disease statistics are simple. A method of enthusiastic contact needs to be established to remember a concrete aim to minimize the measure of the diet of heart disease.

M.Akhiljabbar suggested [19] that heart disease be the only key catalyst for downfall in developed countries and an important supporter of the burden of pain in developing countries. In the country side 30% of ruin is extracted from knowledge from valid sources of India and from coronary disease in the territories of Andhra Pradesh.

V.V.Ramalingam et.al [20] presented an examination of different machine learning algorithms and their performance in relation to cardiovascular disease (CVDs) is a major cause for a large number of deaths in the world over recent decades. There is thus a need for reliable, precise

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

and feasible system for the proper treatment of heart diseases to diagnose such diseases in time.

J.Vijayashree and N.Ch.Sriman Narayana Iyengar have suggested [21] to help a physician predict and diagnose cardiac disease by using computerized support systems. The goal of this review is to discuss the cardiovascular conditions associated with heart and to discuss emerging decision support systems in the field of cardiovascular forecasting and diagnosis, assisted by data mining and hybrid intelligent techniques.

Amita malav and Kalyani kadam suggested [22] that single methods of data mining have fair accuracy in diagnosing and treating heart diseases. However, we can use hybrid data mining techniques to enhance the degree of precision. In order to encourage intelligent decision help, the Heart Disease Prediction (HDP) proposed system guides. A predictive analysis is performed using K-means and ANN data mining techniques on UCI heart disease dataset in the proposed model. The mixture of fuzzy and smooth values is medical data. These data are categorized by their characteristics. The classification is carried out to design of a model using the Artificial Neural Network and k-means algorithm. The focus of our model is based on classifying data better in order to ensure a more accurate diagnosis according to cardiovascular diseases.

C.Beulah Christalin Latha and S.Carolin Jeeva advocated [23] that to investigate an assembly classification approach which was used by combining several classifiers to increase the precision of low grad algorithms. A quantitative analysis approach has been performed to find out how the ensemble technique can be used to improve the precision of cardiovascular disease. A comparative research method has been carried out to find out how the ensemble methodology can be used to enhance accuracy of heart diseases. This paper focuses on improving the accuracy of poor classification algorithms and implementing a medical data set algorithm to demonstrate its usefulness for the early disease prevention. The findings of this study indicate that bagging and boosting strategies are successful to increase the predictive accuracy of the poor category and demonstrate adequate efficiency when the risk of heart disease is established.

3. System Design & Implementation

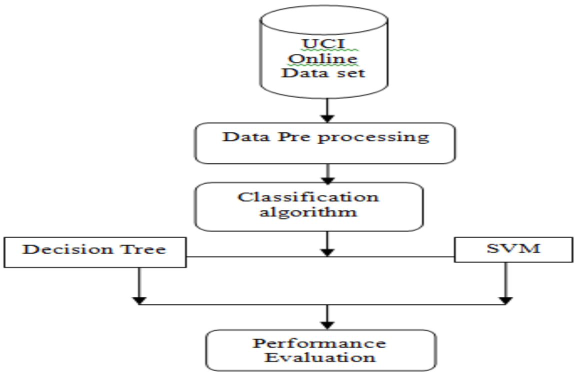

Figure 1. Process Flow of the Hybrid Technique

The collection of different documents is pre-processed for Cardiac disease by. A total of 300 patient records are shown in the dataset and six of them missing. Currently the remaining 290 applications are under way. These six documents were excluded from the data set collection. The attributes of these data set include multi-class vector and binary classificatory. The variable is used for monitoring cardiac disease. If a patient is set to 1, the value is set to zero which indicates a patient without cardiovascular disease. Data is pre-processed by converting medical information into diagnostic values. The previous data showed that 130 reported patients had 1 for cardiovascular disease while the other 150 had 0 for cardiovascular disease failure. The descriptions of the UCI dataset with the attributes used are shown in Table 1**.** The data form and number of values is shown in Table 2.

The 16 features of the data collection defining the personal data of the user will use two features related to age and gender. The remaining other characteristics are significant because they provide valid clinical records. Diagnosis and understanding of the seriousness of cardiac diseases through clinical records. Datasets are clustered on the Decision Tree (DT) parameters and variables. Classifiers are then used in order to estimate the output on each clustered dataset. The best model is described from the above data, based on its low error rate. Through selecting a higher error rate DT cluster, the outcome is also optimized and error management will be calculated for this data collection, to minimize errors.

Decision Trees

A decision tree is a basic representation for classifying instances. It is a Supervised Machine Learning in which the data is continuously divided into a parameter. Entropy reduction is the gain in data. The gain of information is based on the given values and measures the difference between divided entropy and average entropy. The ID3 (Iterative Dichotomiser) decision tree algorithm is based on data gain.

$$Info(D) = -\sum_{i=0}^{m} pijlog2pij \qquad (1)$$

Where, Pi is the probability that an arbitrary tuple in D belongs to class Ci.

$$InfoA(D) = \sum_{j=1}^{n} \frac{|Dj|}{|D|} \times Info(Dj) \qquad (2)$$

$$Gain(A) = Info(D) - InfoA(D) \qquad (3)$$

Where,

- Info (D) is the average amount of information needed to identify the class label of a tuple in D.
- |Dj|/|D| acts as the weight of the jth partition.
- InfoA(D) is the expected information required to classify a tuple from D based on the partitioning by A.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

The attribute A with the highest information gain, Gain (A), is chosen as the splitting attribute at node N.

Support Vector Machine

SVM is a managed machine learning algorithm that can be used for problems with classification and regression. Using the technique known as a kernel trick the data is transformed and on the basis of these transformations it determines an optimal limit between possible outputs. The aim of the supporting vector machine algorithm is to find a hyper plane in an N-dimension (N- the number of attributes) that defines data point.

For Support Vector Classifier (SVC), we use **w**T**x**+*b* where **w** is the weight vector and *b* is the bias.

$$w^T * x + b = 0 \qquad (4)$$

It can be shown that in the hyper plane equation the variables are called **w** and **x,** meaning they are vectors.

Table 1: List of Features Used in this Analysis

Features	Description	Range
Age	Person completed age in years	Numeric
Sex	Person gender (male represented as 1 and female represented as 0)	Nominal
Cp	4 values of chest pain are 1.Typical angina,2.Atypical angina, 3.Non-angina pain and 4.Asymptomatic	Nominal
Trestbps	Blood pressure levels in resting form (in mm/Hg)	Numeric
Chol	Patient cholesterol (in mg/dl)	Numeric
Diabetes	Diabetes<160mg/dl normal	Nominal
Physical activity	The physical habits of human like 1.athelete,2.sedimentry and 3.normal	Numeric
Smoking	Smoking of the patient (false<0.6,true-0.3 to 1)	Nominal
Resting	ECG result while in resting are represented in 3 distinct values: Normal-0,Abnormality in ST T-1and any probability of LV hypertrophy-2	Nominal
Thalach	The make use of of maximum rate of heart	Numeric
Exang	Angina certain by an exercise.(0 depicting ‘no’ and 1 depicting ‘yes’)	Nominal
Oldpeak	Exercise induced ST depression in comparison with the state of rest	Numeric
Slope	ST section considered in terms of slope during peak exercise shown in 3 values:1.unsloping,2.flat and 3.downsloping	Nominal
Ca	Fluoroscopy highlighted major vessels numbered from0 to 3	Numeric

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

Thal	Significance of the heart illustrate through 3 distinctly numbered values. Normal numbered by 3, fixed defect as 6 and reversible defect as 7.	Nominal
Target	Represents the heart disease diagnosis correspond to 5 values, with 0 indicating absence and 1 to 4 representing the presence in different degrees.	Nominal

Table 2: Ranges and Data types of Features

Features	Data Ranges
Age	Numeric[29 to 77; unique=41; mean=54.4; median=56]
Sex	Numeric[0 to 1; unique=2; mean=0.68; median=1]
Cp	Numeric[1 to 4; unique=4; mean=3.16; median=3]
Trestbps	Numeric[94 to 200; unique=50; mean=131.69; median=130]
Chol	Numeric[126 to 564; unique=152; mean=246.69; median=241]
Diabetes	Diabetes<160mg/dl normal
Physical activity	False<0.6; True :1 to 3
Smoking	Smoking of the patient (false<0.6,true-0.3 to 1)
RestECg	Numeric[0 to 2; unique=3; mean=0.99; median=1]
Thalach	Numeric[71 to 202 unique=91; mean=149.61; median=15.3]
Exang	Numeric[0 to 1; unique=2; mean=0.334; median=0.0]
Oldpeak	Numeric[0 to 6.20; unique=40; mean=1.04; median=0.80]
Slope	Numeric[1 to 3; unique=3; mean=1.60; median=56]]
Ca	Categorical [5 levels]
Thal	Categorical [4 levels]
Target	Numeric [0.00 to 4.00;unique=5;mean=0.94;median=0.0]

Pseudo code for Decision Tree

```
Step1  :  Consider training examples, D with positive and negative training set with target class
Step 2 : Description of all the characteristics using for loop
Step 3 :  Do each sample to participate
Step4  : Run the decision tree algorithm
```

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

Step 5 : To identify the features as $f_1, f_2, f_3 \dots f_n$
Step 6 : Complete the number of nodes with its constraints $l_1, l_2, l_3 \dots l_n$
Step 7 : Obtain split the $d_1, d_2, d_3 \dots d_n$

4. Results Discussions

The models predictions have 13 characteristics, and exact modelling techniques are determined. Table 3 measures the precision, classification errors, performance, f-measurement, sensitivity and specificity. The DT-SVM classification cycle achieves the highest accuracy in relation to the current methods. In Figure 2 and figure 3 provides the pictorial representation of overall accuracy and classification error estimated for the dataset used in this research analysis.

Table 3: Outcomes of various models with proposed model

Name of Model	Correctness (%)	Classification error (%)	Precision (%)	F-measure (%)	Sensitivity (%)	Specificity (%)
NB	75.2	23.8	89.8	84.0	78.8	59.8
LR	84.8	14.2	87.9	90.3	94.1	19.5
DL	86.89	11.9	89.9	91.8	94.6	32.8
DT	84.6	14.6	85.7	90.9	97.9	0.0
RF	85.8	12.8	86.8	91.7	97.9	9.4
SVM	85.7	13.1	85.6	91.8	99.5	0.0
DT-SVM	88.3	11.1	89.8	89	91.6	81.8

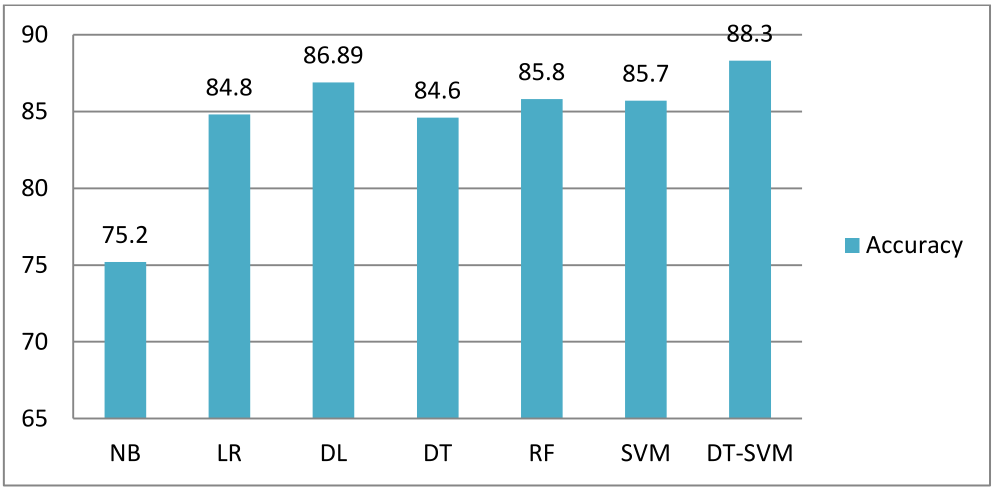

Figure 2: Overall accuracy of the given dataset by using various models

It is clearly proved from the Figure 2 that the overall accuracy of the DT-SVM method is 88.3% and is very high when compared to all other methods. Similarly in Figure 3

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

classification error rate of the DT-SVM method very less when compared to all other method considered in this research analysis.

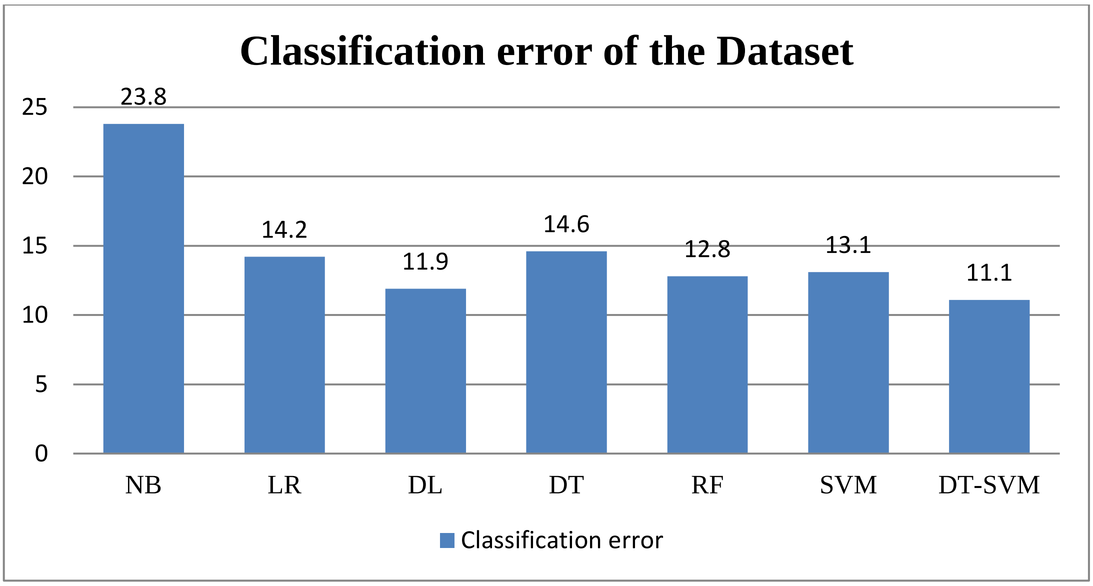

Figure 3: overall classification error of given dataset

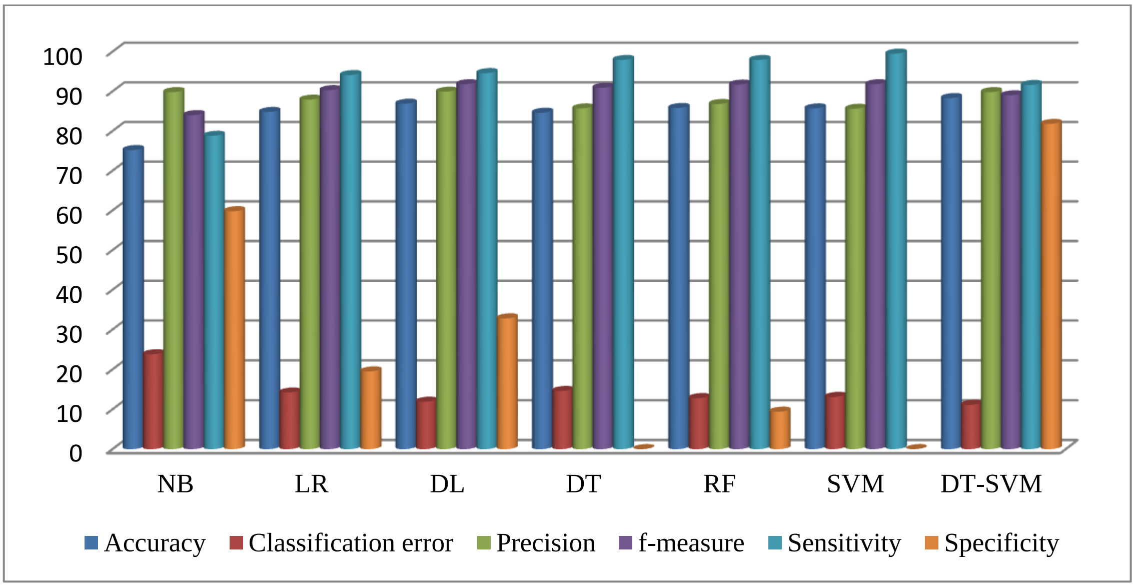

Figure 4: Feature comparison of the given dataset by using various models

In order to compare current models output with that of the proposed model, benchmarking is required. This approach is used to determine how the proposed system is optimally accurate. Accuracy with the number of selected functions and results obtained from the model is

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

determined. The range of functionality to be used by DT-SVM is without limit. The best results are found for every function selected for this model.

5. Conclusion

The decision support system have been used to collect raw health data from hospital can help to save human lives and avoid heart complications in the longer term. The methods used for processing raw data and for discerning heart disease were used in this study. In the medical field, prediction of cardiopathy is challenging and critical. Nonetheless, the death rate may be greatly decreased if the disease is diagnosed early in life and treatment measures are taken as soon as possible. An extension of this analysis is highly desirable for investigating real world data sets instead of theoretical methods and simulations. In combining features of DT and SVM, a proposed hybrid DT-SVM solution is used. DT-SVM showed that predicting heart disease was reliable. The potential direction of this work can be conducted using various combinations of machine learning technologies to enhance prediction. In addition, a broader understanding of the essential features to improve predictive cardiomyopathies can be established using the modern feature selection methods.

6. References

[1]. Pasquale Arpaia, Carlo Manna, Giuseppe Montenero, and Giovanni D’Addio “In-Time Prognosis Based on Swarm Intelligence for Home-Care Monitoring: A Case Study on Pul monary Disease” IEEE Sensors Journal, VOL. 12(3),MARCH 2012,pp 692-698.

[2]. Booma Devi Sekar, Ming Chui Dong, Jun Shi, and Xiang Yang Hu “Fused Hierarchical Neural Networks for Cardiovascular Disease Diagnosis” IEEE Sensors Journal, VOL. 12, NO. 3, MARCH 2012 pp 644- 650.

[3]. Daniel Bia Santana, Yanina A. Z´ocalo, and Ricardo L. Armentano, Senior Member, IEEE “Integrated e-Health Approach Based on Vascular Ultrasound and Pulse Wave Analysis for Asymptomatic Atherosclerosis Detection and Cardiovascular Risk Stratification in the Community” IEEE Transactions On Information Technology In Biomedicine, VOL. 16, NO. 2, MARCH 2012 pp 287-294.

[4].Kersten Petersen, Melanie Ganz, Peter Mysling, Mads Nielsen, Lene Lillemark, Alessandro Crimi, and Sami S. Brandt “A Bayesian Framework for Automated Cardiovascular Risk Scoring on Standard Lumbar Radiographs” IEEE Transactions On Medical Imaging, VOL. 31, NO. 3, MARCH 2012 pp 663-676.

[5].SurajitBagchi, Subhabrata Sengupta, and Sanjoy Mondal,” Development and Characterization of Carbonic Anhydrase-Based CO2 Biosensor for Primary Diagnosis of Respiratory Health “IEEE Sensors Journal, VOL. 17, NO. 5, MARCH 1, 2017 pp 1384-1390.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

[6]. Yi-Ting Cheng, Yu-Feng Lin, Kuo-Hwa Chiang, and Vincent S. Tseng "Mining Sequential Risk Patterns From Large-Scale Clinical Databases for Early Assessment of Chronic Diseases: A Case Study on Chronic Obstructive Pulmonary Disease" IEEE Journal Of Biomedical And Health Informatics, VOL. 21, NO. 2, MARCH 2017 pp303-311.

[7]. Shubpreet Kaur and Dr. R.K.Bawa " Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System" International Journal of Energy, Information and Communications ,VOL..6, Issue 4 (2015), pp.17-34.

[8]. Mario Merone, Claudio Pedone, Giuseppe Capasso, Raffaele Antonelli Incalzi, and Paolo Soda "A Decision Support System for Tele-Monitoring COPD-Related Worrisome Events" IEEE Journal of Biomedical And Health Informatics, VOL. 21, NO. 2, MARCH 2017 pp 296-302.

[9] .Cheng YT, Lin YF, Chiang KH, Tseng VS, "Mining Sequential Risk Patterns From Large-Scale Clinical Databases for Early Assessment of Chronic Diseases: A Case Study on Chronic Obstructive Pulmonary Disease", "IEEE Journal of Biomed Health Information", VOL.2, March 2017, pp. 303-311.

[10]. Sudha Ram, Wenli Zhang, Max Williams, Yolande Pengetnze, "Predicting Asthma-Related Emergency Department Visits Using Big Data", IEEE Journal Of Biomedical And Health Informatics, VOL. 19, Issue:4, July 2015, PP.1216-1223.

[11]. J. Thomas ; R Theresa Princy, "Human heart disease prediction system using data mining techniques", "IEEE international Conference on Circuit, Power and Computing Technologies (ICCPCT)", 18-19 March 2016.

[12]. Eman AbuKhousa, Piers Campbell "Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems" International Conference on Innovations in Information Technology (IIT), 2012 pp 267-272

[13] .Chaitrali S. Dangare and Sulabha S.Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications , Vol.47, No. 10, pp.0975-888, 2012.

[14]. N. Deepika and K... Chandrashekar, "Association rule for classification of Heart Attack Patients", International Journal of Advanced Engineering Science and Technologies, Vol.11, No.2, pp253-257, 2011.

[15]. Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol.8, No.1, 2007

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.239

[16] .Asghar, S. “Automated Data Mining Techniques: A Critical Literature Review” 978-0-7695-3595-1, 75 – 79, IEEE, 2009.

[17]. Parisa Naraei, Abdolreza Abhari and Alireza Sadeghian,”Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data”, IEEE, 2016.

[18] .DeepaliChandna “Diagnosis of Heart Disease Using Data Mining Algorithm”, IEEE Conf. on International Journal of Computer Science and Information Technologies, 2015, pp 1678-1680

[19] .M.Akhiljabbara“Heart Disease Prediction System using Associative Classification and Genetic Algorithm”IEEE, 2012.

[20] .V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja,” Heart disease prediction using machine learning techniques : a survey” in International Journal of Engineering & Technology(IJET) ,2018.

[21]. J.Vijayashree, N.Ch.SrimanNarayanaIyengar “Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review” in International Journal of Bio-Science and Bio-Technology (IJBS), VOL.8.Issue:4 August 2016.

[22] .Amita malav ,Kalyana kadam,” A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K-means” in International Journal of Pure and Applied Mathematics (IJPAM),2018.

[23]. C.BeulahChristalinLatha, S.CarolinJeeva,” Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”,Elsevier journal,2019.