International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 22 (2018) pp. 15786-15792

Heart Disease Prediction on Continuous Time Series Data with Entropy Feature Selection and DWT Processing

[1]Veena N, [2]Dr.Anitha N

[1]Research Scholar(VTU), Information Science and Engineering,
B M S Institute of Technology and Management, Avalahalli, Yelahanka, Bangalore-560064, India.

[2]Professor, Information Science and Engineering, East Point College of Engineering and Technology,
Virgonagar(post), Bidarhalli, Bangalore-560049, India.

Abstract

Heart Disease is major death causing disease in many parts of world as per world health organization reports. With food habits and stressful living conditions heart disease occurs even in younger age groups. To prevent any death risk, timely prediction and treatment is necessary. In this work, features for prediction of heart disease for data mining are explored using feature selection and wavelet feature extraction. The performance of different machine learning classifier for prediction of heart disease is measured and compared.

I. INTRODUCTION

Every year 17.5 million deaths are reported due to cardiovascular diseases worldwide. Almost 80% of them are due to heart attack and stroke. India has fastest growing number of cardio vascular disease patients in recent years due to drastic shift in quality of life and food habits. India has more than 30 million heart disease patients and over two lakh open heart surgeries are performed each year. A more worrying concern is that number of patients requiring coronary interventions is rising at alarming rate of 20% to 30% every year.

Many of heart related deaths could have been avoided if it has been diagnosed in time and effective treatment procedures have been started in advance. Heart disease diagnosis is complicated and depends on doctor expertise. Due to diagnosis cost and time involved, patients hardly go through regular diagnosis procedure. This necessitates use of low cost automated diagnosis system which can save cost and time for the patients. Use of data mining is explored in this work for providing a low cost automated diagnosis system for patients.

Huge amount of data is collected from patients by health care institutions nowadays. Mining on these data, various inferences on disease prediction and categorization can be done. The problem in mining on these huge data is selection of attributes relevant for classification of the disease. Once the features are selected, the classification model is built using various algorithms like KNN, neural network, naïve baiyes, SVM etc.

In this paper, a entropy based analysis is done on correlation between attributes and the heart disease risk. Based on the correlation analysis features are selected and with the selected features time dynamics on continuous time period is captured using wavelet based statistical features. With the final features set, three machine learning models using KNN, SVM and ANN are trained and their classification accuracy is compared. Most of existing works on heart disease diagnosis is based on attributes values collected on particular time instant. But this way of prediction is not always accurate. Say weight parameter influences heart attack, but say the patient is on reducing weight and the risk is low due to reducing weight, this scenario cannot be modeled on data collected at a particular time snap and used for classification. This necessities collection of data on a continuous interval of time and the time on a window time interval as such should be used for classification. By this way the accuracy of the classification increases. In this work, continuous data modeling is used for increased accuracy.

II. RELATED WORK

In this section the previous works on machine learning for heart disease diagnosis and classification is explored.

In [1] authors analyzed the performance of five classifiers. The classifier were Bayes Net, SMa, KStar, MLP and J48. From their analysis Bayes Net and SMa classifiers are the optimum among the all five classifiers. Their analysis is based on data set collected at particular time and not on continuous time series data.

In [2] author detailed different classification methods for heart disease prediction. They studied decision trees, neural networks and naïve baiyes classifier for prediction. The dataset attribute selection was not given attention in this work.

In [3] authors studies the use of knowledge discovery process to analyze Stroke using ANN and SVM classifiers. Feature selection was done using Cramer's V test to select the attributes. Around 80% accuracy was achieved in this method. One of important inference in their analysis is feature selection gives better accuracy.

In [4] authors proposed a heart disease classification system. Features are extracted from ECG signals and used for classification. Higher-Order Statistics bispectrum and cumulant features are extracted from ECG beat. The features extracted are applied to Principal Component Analysis (PCA) to reduce the dimensions. Then PCA coefficients are ranked

using multiple ranking methods. KNN and Decision tree models are built using ranked features to get the highest accuracy. The classification cannot predict initial stages of disease.

In [5] authors implemented a expert system using decision tree and KNN classifier to predict disease. They used Pima Indians Diabetes Database to predict the diabetes. The disadvantage in this work is that it is based on snapshot data and not on continuous time series data.

In [6] authors proposed a decision support system for disease classification with more focus on feature selection. Probabilistic Principal Component Analysis (PPCA) was used for feature extraction. It extracted features with high covariance and feature dimension is also reduced in this work. On the dimension reduced data, RBF neural network is built to predict the disease. With usage of feature selection accuracy improved, but the dataset is not continuous and it is only snapshot.

In [7] authors proposed a risk model for Heart disease. This model can predict 1-year or more survival for Heart failure diagnosed patients. From the unstructured medical records, certain attributes are extracted using search and then multinominal Naïve Bayes (NB) is used for classification. The approach does not address missing value problem. Feature selection is not given attention.

In [8] authors analyzed heart disease prediction problem using six machine learning techniques. Thirteen distinct attributes from dataset is taken for analysis. StatLog heart disease dataset from UCI machine learning repository is taken and classifier models built on that dataset to predict heart disease. But the work does not give attention to feature set selection and continuous data.

In [9] authors proposed a prediction system for coronary heart disease. It used narrative medical histories and extracted features from it for training and classification. From the unstructured text input, features are extracted using natural language processing techniques and from it PCA , mutual information filter employed to reduce feature dimension size. SVM classifier is then built to classify. The system accuracy depends on narration ability of patients and when missing information in narration, the accuracy is very low.

In [10] authors used decision tree classifier to diagnose Heart Arrhythmia. Discrete wavelet transform (DWT) is done on ECG signals to get coefficients and from it statistical features are extracted and used as features. The accuracy of the method is not satisfactory as no feature dimensionality reduction procedures are followed. Also data is snap shot based and not on continuous based.

In [11] authors proposed a combined classifier to classify heart diseases. It used KNN and genetic algorithm as a combination. Genetic algorithm removed redundant and irrelevant attributes, and for ranking the attributes which contribute more towards classification. Low ranked attributes are removed and KNN model is built on it for classification. By using Genetic search classification accuracy is improved in this approach.

In [12] authors devised a heart disease classification using ECG signals. Signal Denoising is done using wavelet filters and on the filtered signal, Discrete wavelet transform (DWT) based features are extracted. The method is only for data collected at a particular snapshot.

In [13], author proposed a method based on selecting best features with random forest algorithm. After selected features are extracted, different machine learning techniques are applied to predict heart disease. They were able to achieve 75% accuracy using this method.

In [14] authors proposed a machine learning classification for heart diseases. They used logistic regression, support vector machines and neural networks for classification. Cleveland Heart Disease Dataset was used for analysis. But the accuracy is not good enough in this approach and it is not on continuous data.

III. HEART DISEASE DIAGNOSIS

A Heart disease diagnosis system is proposed in this paper. The proposed approach consists of three parts.

1. Feature Selection
2. Preprocessing
3. Machine Learning Classification

In feature selection, the necessary features for classification of heart disease are selected and with the selected features machine learning classifier models are built for classification.

Feature Selection

Feature Selection is done to identify relevant attributes which influences the output. It is done mainly to reduce the training time and build a more accurate model.

The feature selection is based on calculation of symmetric uncertainty between attributes and target class. Once the symmetric uncertainty is calculated between each attribute and the target class , the attributes who symmetric uncertainty value less than a threshold value are dropped as irrelevant features and rest all features are selected for classification.

Symmetric uncertainty is calculated as

$$SU(X,Y) = \frac{2 \times MI(X,Y)}{H(X) + H(Y)}$$

Where

The entropy of a attribute X is given as H(X)

The entropy of a class variable Y is given as H(Y)

The mutual information between X and Y is given as MI.

Mutual Information (MI) is calculated as

$$MI(X,Y) = \sum_{x}\sum_{y} p(x,y)\log\frac{p(x,y)}{p(x)*p(y)}$$

Using Shannon entropy the mutual information between two variables is calculated as

$$MI(X,Y) = H(X,Y) - H(X|Y) - H(Y|X)$$

Where

$$H(X) = -\int p(x)\log(p(x))dx$$

The feature selection procedure is given below

Input : Data set with N attributes and the class column Y , Thres

Output:SelFeatures

Suvect=[];

For i=1: N

Suvect ← Suvect + SU(Xi,Y)

End

SelFeatures=[]

For i=1:N

If SuVect(i)>Thres

SelFeatures= SelFeatures + i;

End

End

Preprocessing

From the data collected for the patient over a period of time (say collected every day), the selected features alone are taken. From the selected features time variants attributes are arranged as time series , The time series data is segmented to smaller overlapped sliding windows .The window is a configurable parameter. Each sliding window is passed through Fast Fourier transformation.

The DFT can be defined as

$$X(P) = \sum_{t=0}^{T-1} x(t)W_T^{tp} \Leftrightarrow x(t) = \frac{1}{T}\sum_{p=0}^{T-1} X(P)W_T^{-tp},$$
$$where \quad W = c^{-j2\Pi/T}$$

The DFT can be presented as a discrete-time Fourier transform of a cyclic signal with period T as given below

$$x = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(T-1) \end{bmatrix}, \quad X = \begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(T-1) \end{bmatrix}$$

$$W = [W_T^{pt}] = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & W_T & \dots & W_T^{T-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W_T^{T-1} & \dots & W_T^{(T-1)(T-1)} \end{bmatrix}$$

By applying the Fast Fourier transformation, the coefficients of time series data over frequency domain is captured. On the coefficients resulting from Fast Fourier transformation, following statistical features are extracted.

Maximum Value

Minimum Value

Mean

Standard Deviation

Feature name	Formula
Maximum value	$X_{Max} = Max[x_n]$
Minimum value	$X_{Min} = Min[x_n]$
Mean	$X_{Mean} = \frac{1}{n}\sum_{1}^{n} x_i$
Standard Deviation	$X_{SD} = \sqrt{\sum_{n=1}^{N}(x_n - AM)\frac{2}{n-1}}$

The data set with selected features is now transformed as every selected time variant feature is replaced by its statistical features.

Classification

Three classifier models are built on preprocessed training data set. Following models are built

1. LSVM (Linear Support Vector Machine)
2. KNN (K-Nearest Neighbor)
3. ANN

Linear Support Vector Machine

Linear support vector machine is a supervised machine learning technique. It is based on statistical learning theory. Two class problems are solved better using Linear Support Vector Machine. The dataset is mapped to high dimensional space and a hyperplane is constructed to split the classes. The hyperplane is constructed in order to maximize the distance between the plane and the support vectors.

For a training data with n points (x1,y1),(x2,y2),…(xn,yn) L-SVM finds the optimal hyperplane by using the following rules

$$y_i[(wx_i) + w_0] = 1 - \xi_i, i = 1, ..., m,$$

$$1/2\|w\|^2 + \frac{c}{2}\sum_{i=1}^{m}\xi_i^2$$

Problem solved by L-SVM is formulated as

$$(w, b, \alpha, \xi) = 1/2\|w\|^2 + \frac{c}{2}\sum_{i=1}^{m}\xi_i^2$$
$$-\sum_{i=1}^{m}\alpha_i\{y_i[(wx_i) + w_0] - 1 + \xi_i\}$$

Artificial Neural Network

ANN is a supervised machine learning technique. It users a set of neurons interconnected in layers in its core to remember the training pattern. A three layer perceptron is used with back propagation learning as in Fig 1.

ANN is designed with following considerations

The Number of input layer neurons is same as that of Number of attributes in processed dataset

The Number of hidden layer neuron is usually twice the Number of input layer neuron

The number of output class is the Number of output layer neuron.

Transfer function used at neuron = tanh.

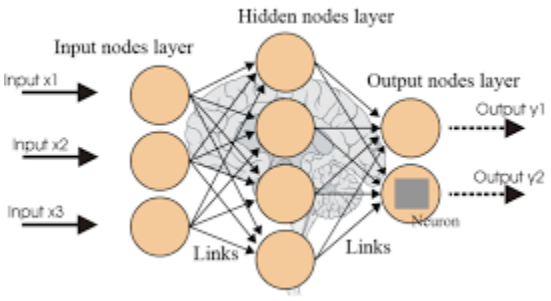

Figure 1: Artificial Neural Network with hidden Layers

K Nearest Neighbor

This is the simplest of all classifier. In this the output class for a test is decided based on the majority of the neighbor's output class as in Fig 2.

The neighbor is calculated based on Euclidean distance between data points.

Say dataset p1 at (x1, y1) and dataset p2 at (x2, y2), Euclidean distance is $\sqrt{((x1 - x2)^2 + (y1 - y2)^2)}$.

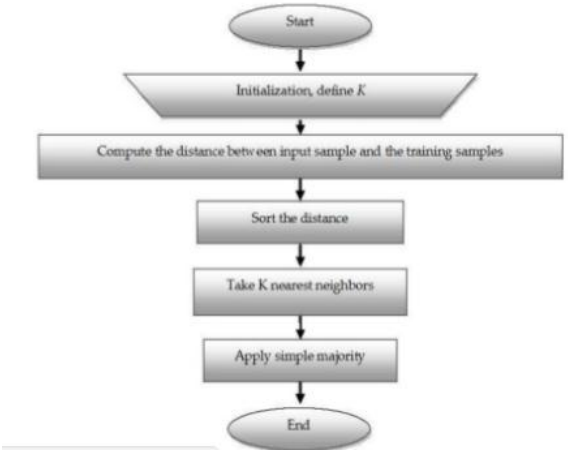

Figure 2: Steps of K nearest neighbors Algorithm

The overall process flow of the proposed solution is shown in Fig 3.

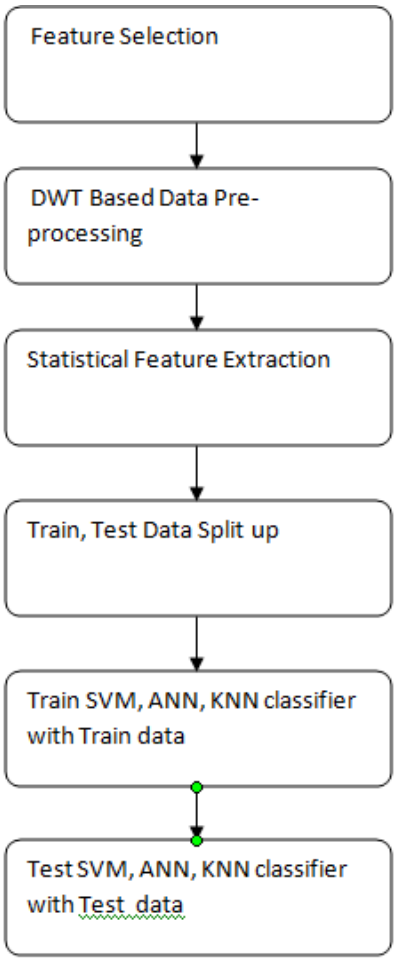

Figure 3: Overall process of proposed work

IV. RESULTS

The proposed solution was tested on South Africa Heart Disease Dataset [15]. The dataset consist of heart disease samples in high risk populations. The data set is collected from patients who have undergone chronic health disease tests , blood pressure reduction treatment and other programs devised for reduction of health risks

After executing Symmetric Uncertainty based feature selected following attributes are selected form the dataset.

- sbp systolic blood pressure
- tobacco cumulative tobacco (kg)
- ldl low densiity lipoprotein cholesterol
- adiposity
- famhist family history of heart disease (Present, Absent)
- typea type-A behavior
- obesity
- alcohol current alcohol consumption
- age age at onset
- chd response, coronary heart disease

From this 6,9 are time invariants and rest are time variant data. Fast Fourier transformation is done on time variant and 4 statistical features are extracted for each time variant.

The total number of attributes in the data set after pre processing is 8*4 + 2 = 34.

The dataset is split to 80:20 ratio , where 80% is used for training and 20 % is used for testing.

Table 1: L-SVM is trained with following parameters

Degree	3
Gamma	0.5

Table 2: The ANN is trained with following parameters

Layers	3
Input Neurons	34
Hidden Layer Neuron	69
Output Neuron	2

Symmetric uncertainty value in percentage for all relevant attributes selected as in Fig 4

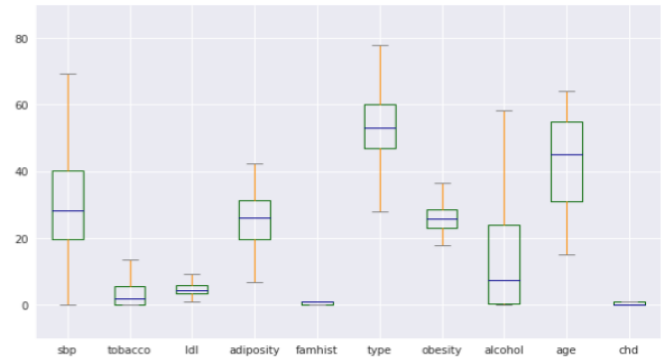

Figure 4: Symmetric uncertainty value in percentage for all relevant attributes

For all the 3 classifiers, following performance metrics are measured.

1. Accuracy = TP + TN / (TP + FP + FN + TN)
2. Precision = TP / (TP+FP)
3. Recall = TP / (TP + FN)

Where

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Accuracy, Precision & recall is measured for each of three classifiers and result is given Fig 5.

Figure 5: Accuracy, Precision & recall is measured for SVM, KNN, ANN and their result

The percentage of improvement in accuracy with and without DWT based preprocessing is measured and plotted as in Fig 6

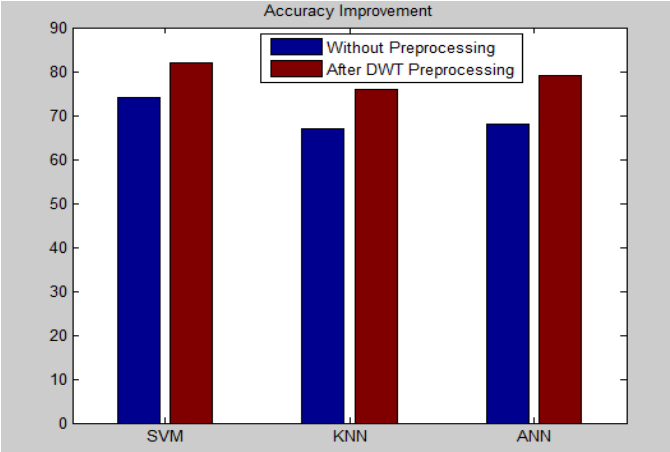

Figure 6: Improvement in accuracy with and without DWT based preprocessing is measured

DWT preprocessing is able to increase the classification accuracy by average 8%.

The features selected was varied by varying the threshold and the accuracy of classification in all three classifier for different values of threshold is given below

The accuracy with and without pre-processing for SVM classifier is measures for various thresholds for feature selection and plotted as in Fig 7.

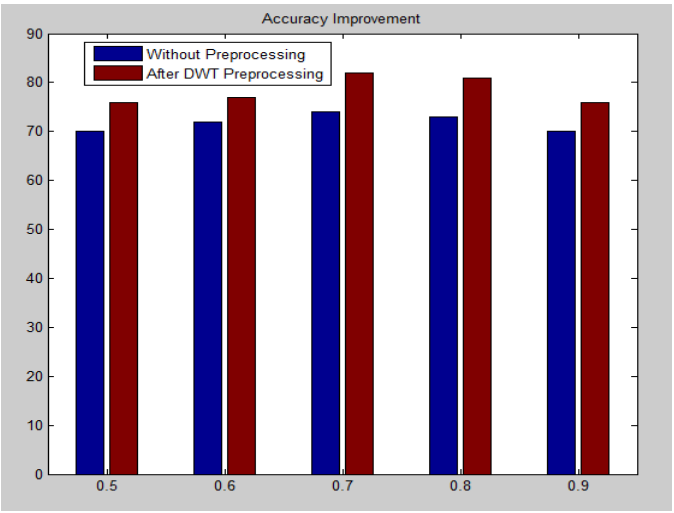

Figure 7: accuracy with and without pre-processing for SVM classifier is measures for various thresholds for feature selection

From the results, it can be seen that maximum accuracy is achieved at 0.7.

The accuracy with and without pre-processing for ANN classifier is measured for various thresholds for feature selection and plotted as in Fig 8.

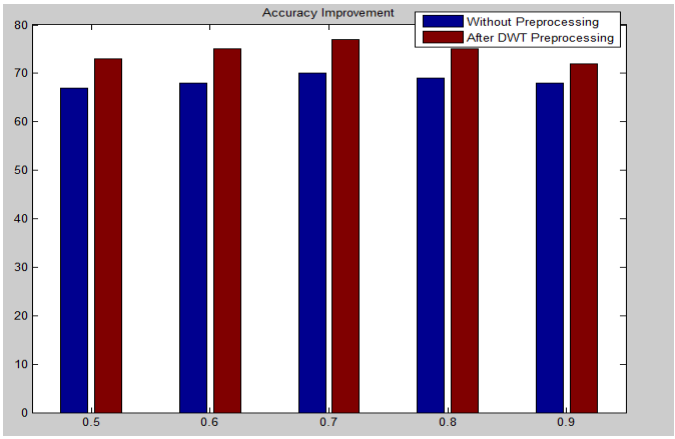

Figure 8: accuracy with and without pre-processing for ANN classifier is measured for various thresholds for feature selection

From the results, it can be seen that maximum accuracy is achieved at 0.7.

The accuracy with and without pre-processing for KNN classifier is measured for various thresholds for feature selection and plotted as in Fig 9.

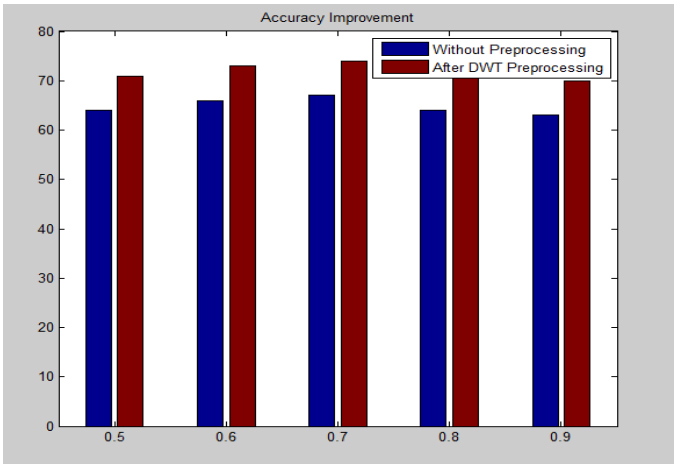

Figure 9: accuracy with and without pre-processing for KNN classifier is measured for various thresholds for feature selection

From the results, it can be seen that maximum accuracy is achieved at 0.7

The accuracy is measured for varying interval of window (k value) in DFT during data preprocessing and for all three classifier is shown in Fig 10.

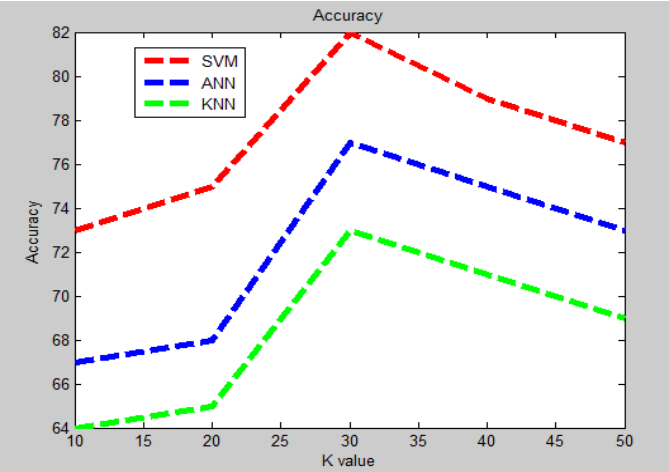

Figure 10: Accuracy is measured for varying interval of window (k value) in DFT during data preprocessing and for all three classifiers

The best value of accuracy is achieved at window size of 30.

V. CONCLUSION

In this work, a heart disease diagnosis method was proposed. With the proposed method features are selected and pre-processing done on training data set. On the preprocessed dataset three different classifier models were trained and tested for accuracy. From the result all the classifiers were found to return more that 80% accuracy and SVM performs best with 87%. Compared to individual classifier, ensemble classifier was found to give better accuracy in many classifications, so future work would be on usage of ensemble classification on the preprocessed data set.

REFERENCES

[1] M. Sultana, A. Haider, and M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016, 2017

[2] M. Gandhi, “Predictions in Heart Disease Using Techniques of Data Mining,” Int. Conf. Futur. trend Comput. Anal. Knowl. Manag., 2015.

[3] C. Colak, E. Karaman, and M. G. Turtay, “Application of knowledge discovery process on the prediction of stroke,” Comput. Methods Programs Biomed., vol. 119, no. 3, pp. 181–185, 2015.

[4] U. R. Acharya et al., “Application of higher-order spectra for the characterization of Coronary artery disease using electrocardiogram signals,” Biomed. Signal Process. Control, vol. 31, pp. 31–43, 2017.

[5] E. K. Hashi, M. S. U. Zaman, and M. R. Hasan, “An expert clinical decision support system to predict disease using classification techniques,” 2017 Int. Conf. Electr. Comput. Commun. Eng., pp. 396–400, 2017..

[6] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, “Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis,” Phys. A Stat. Mech. its Appl., vol. 482, pp. 796–807, 2017.

[7] M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, “Identification of Heart Failure by Using Unstructured Data of Cardiac Patients,” 2016 45th Int. Conf. Parallel Process. Work., pp. 426–431, 2016..

[8] A. K. Dwivedi, “Performance evaluation of different machine learning techniques for prediction of heart disease,” Neural Comput. Appl., pp. 1–9, 2016.

[9] K. Buchan, M. Filannino, and Ö. Uzuner, “Automatic prediction of coronary artery disease from clinical narratives,” J. Biomed. Inform., vol. 72, pp. 23–32, 2017.

[10] E. Alickovic, A. Subasi, "Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier", *Patient Facing System J Med Syst*, vol. 40, pp. 108-120, 2016.

[11] M. Jabbar, B. Deekshatulua, P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm", *Procedia Tech.*, vol. 10, pp. 85-94, 2013.

[12] S. U. Kumar and H. H. Inbarani, “Neighborhood rough set based ECG signal classification for diagnosis of cardiac diseases,” Soft Computing, vol. 21, no. 16, pp. 4721–4733, 2017.

[13] A. Mustaqeem, S. M. Anwar, M. Majid, and A. R. Khan, “Wrapper method for feature selection to classify cardiac arrhythmia,” in Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3656–3659, July 2017

[14] D. Khanna, R. Sahu, V. Baths, and B. Deshpande, “Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease,” International Journal of Machine Learning and Computing, vol. 5, no. 5, pp. 414–419, 2015.

[15] South Africa Heart Disease Dataset Source: https://web.stanford.edu/~hastie/ElemStatLearn//data.html (https://web.stanford.edu/~hastie/ElemStatLearn//data.html) https://www.openml.org/d/1498 (https://www.openml.org/d/1498)