Study and Development of Novel Feature Selection Framework for Heart Disease Prediction

T. John Peter*, K. Somasundaram**

*Dept. of IT, KCG College of Technology, Chennai
**Dept. of CSE Jeya College of Engineering, Chennai

Abstract- Heart disease prediction is designed to support clinicians in their diagnosis. We proposed a method for classifying the heart disease data. The patient's record is predicted to find if they have symptoms of heart disease through Data mining. It is essential to find the best fit classification algorithm that has greater accuracy on classification in the case of heart disease prediction. Since the data is huge attribute selection method used for reducing the dataset. Then the reduced data is given to the classification .In the Investigation, the hybrid attribute selection method combining CFS and Filter Subset Evaluation gives better accuracy for classification. We also propose a new feature selection method algorithm which is the hybrid method combining CFS and Bayes Theorem. The proposed algorithm provides better accuracy compared to the traditional algorithm and the hybrid Algorithm CFS+FilterSubsetEval.

Index Terms- Data mining, Feature Selection and Classification.

I. INTRODUCTION

THe Heart Disease Data Prediction is designed to support clinicians in their diagnosis for heart disease prediction. They typically work through an analysis of medical data and a knowledge base of clinical expertise. The quality of medical diagnostic decisions for heart disease can be increased by improvements to these Predicting systems. Data mining provides a way to get the information buried in the data. Here we propose the hybrid Algorithm combining the best Feature Selection methods for Classification. Since, We have large collections of data which consumes more time for classification. The patient's record is classified and predicted if they have the symptoms of heart disease. It is essential to find the best fit algorithm that has greater accuracy on classification in the case of heart disease classification.

The large number of data can be reduced that using hybrid attributes selection methods. In order to find the best two algorithms of attribute selection method, the attribute selection method which gives higher accuracy after removing attributes for both classification and clustering are identified and combined to form the hybrid attribute selection method. Then the reduced data are fed into sequence of classifiers classified to attain better accuracy.

A. Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. By removing most irrelevant and redundant features from the data, feature selection helps improve the performance of learning models.

B. Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

A learning classifier is able to learn based on a sample. The dataset used for training consists of information x and y for each data-point, where x denotes what is generally a vector of observed characteristics for the data-item and y denotes a group-label. The label y can take only a finite number of values.

C. Organization of the Paper

The Second chapter is a literature survey on heart disease prediction, and Feature selection. The third chapter contains the proposed methodology which includes the Proposed Hybrid Feature Selection Algorithm. The Fourth chapter shows the Results and Discussion. The Fifth chapter concludes the paper with Further Directions.

II. LITERATURE REVIEW

A novel technique to develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) was proposed by Heon Gyu Lee et al[11]. Statistical and classification techniques were utilized to develop the multi-parametric feature of HRV. Besides, they have assessed the linear and the non-linear properties of HRV for three recumbent positions, to be precise the supine, left lateral and right lateral position. Numerous experiments were conducted by them on linear and nonlinear characteristics of HRV indices to assess several classifiers, e.g., Bayesian classifiers [11], CMAR (Classification based on Multiple Association Rules) [19], C4.5 (Decision Tree) [24] and SVM (Support Vector Machine) [4]. SVM surmounted the other classifiers.

A model Intelligent Heart Disease Prediction System (IHDPS) built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Neural Network was proposed by Sellappan Palaniappan et al. [26]. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease. IHDPS subsists well being web-based, user-friendly, scalable, reliable and expandable.

The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. [22].

Experiments were carried out on a sample database of patients' records. The Neural Network is tested and trained with 13 input variables such as Age, Blood Pressure, Angiography's report and the like. The supervised network has been recommended for diagnosis of heart diseases. Training was carried out with the aid of back propagation algorithm. Whenever unknown data was fed by the doctor, the system identified the unknown data from comparisons with the trained data and generated a list of probable diseases that the patient is vulnerable to.

The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [3]. The assessed data set encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing. Three constraints were introduced to decrease the number of patterns. First one necessitates the attributes to appear on only one side of the rule. The second one segregates attributes into uninteresting groups. The ultimate constraint restricts the number of attributes in a rule. Experiments illustrated that the constraints reduced the number of discovered rules remarkably besides decreasing the running time. Two groups of rules envisaged the presence or absence of heart disease in four specific heart arteries. Data mining methods may aid the clinicians in the prediction of the survival of patients and in the adaptation of the practices consequently.

The work of Franck Le Duff et al. [9] might be executed for each medical procedure or medical problem and it would be feasible to build a decision tree rapidly with the data of a service or a physician. Comparison of traditional analysis and data mining analysis illustrated the contribution of the data mining method in the sorting of variables and concluded the significance or the effect of the data and variables on the condition of the study.The main drawback of the process was knowledge acquisition and the need to collect adequate data to create an appropriate model.

A novel heuristic for efficient computation of sparse kernel in SUPANOVA was proposed by Boleslaw Szymanski et al. [25]. It was applied to a benchmark Boston housing market dataset and to socially significant issue of enhancing the detection of heart diseases in the population with the aid of a novel, non-invasive measurement of the heart activities on basis of magnetic field generated by the human heart. 83.7% predictions on the results were correct thereby outperforming the results obtained through Support Vector Machine and equivalent kernels.The spline kernel yielded equally good results on the benchmark Boston housing market dataset. In [17] Latha Parthiban et al. projected an approach on basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. The CANFIS model diagnosed the presence of disease by merging the neural network adaptive capabilities and the fuzzy logic qualitative approach and further integrating with genetic algorithm. On the basis of the training performances and classification accuracies, the performances of the CANFIS model were evaluated. The CANFIS model is promising in the prediction of the heart disease as illustrated by the results.

In [29] Kiyong Noh et al. put forth a classification method for the extraction of multi-parametric features by assessing HRV from ECG, data preprocessing and heart disease pattern. The efficient FP-growth method was the basis of this method which is an associative. They presented a rule cohesion measure that allows a strong push of pruning patterns in the pattern generating process as the volume of patterns created could possibly be huge. The multiple rules and pruning, biased confidence (or cohesion measure) and dataset consisting of 670 participants, distributed into two groups, namely normal people and patients with coronary artery disease, were employed to carry out the experiment for the associative classifier.

In Lei Yu and Huan Liu [19] introduced a novel concept, predominant correlation, and proposed a fastfilter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. The efficiency and effectiveness of their method is demonstrated through extensive comparisons with other methods using real world data of high dimensionality.

A methodology for comparing classification methods through the assessment of model stability and validity in variable selection was proposed by J. Shreve , H.Schneider , O. Soysal[16]. This study provides a systematic design for comparing the performance of six classification methods using Monte Carlo simulations and illustrates that the variable selection process is integral in comparing methodologies to ensure minimal bias, enhanced stability, and optimize performance. They quantify the variable selection bias and show that, for sufficiently large samples, this bias is minimized so that methods can be compared.

Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels proposed by Tinghua Wanga, Houkuan Huang, Shengfeng Tian , Jianfeng Xu [7]. This work focused on effective feature selection method for support vector machine (SVM). Unlike the traditional combinatorial searching method, feature selection is translated into the model selection of SVM . The basic idea of this method is to tune the hyperparameters of the Gaussian Automatic Relevance Determination (ARD) kernels via optimization of kernel polarization, and then to rank all features in decreasing order of importance so that more relevant features can be identified.

Feature selection for Bayesian network classifiers using the MDL-FS score proposed by Madalina M. Drugan , Marco A. Wiering[8] .They propose a new definition of the concept of redundancy in noisy data. They show that the MDL-FS function serves to identify redundancy at different levels and is able to eliminate redundant features from different types of classifier.

Sarah Ashoori and Shahriar Mohammadi[6] proposed a Comparison of failure prediction models based on feature selection Technique. A huge amount of information about the corporations that derived from financial reports could be used to determine the failure of companies, but it needs much time and human resources. For Selection of financial variables or features ,Two prediction models would be compared with each other in 3 stages. These models are neural networks that named "MultiLayer Perceptron". One of these models is trained with original dataset and the other one is trained with a dataset contained the selection of features of original data set.

A decision rule-based method for feature selection in predictive data mining proposed by Patricia E.N. Lutu Andries P. Engelbrecht[2]. This method incorporates domain specific definitions of high, medium and low correlations techniques. The

ISSN 2250-3153

proposed algorithm conducts a heuristic search for the most relevant features for the prediction task.

Tina Tirelli , Daniela Pessani proposed the importance of feature selection in decision-tree and artificial-neural-network ecological applications. [15]. In this work,They use four different feature selection methods (χ2, Information Gain, Gain Ratio, and Symmetrical Uncertainty) and evaluate their effectiveness in preprocessing the input data to be used for inducing artificial neural networks (ANNs) and decision trees (DTs).

Fatemeh Amiri , MohammadMahdiRezaeiYousefi , CaroLucas , AzadehShakery , NasserYazdani proposed Mutual information based feature selection for intrusion detection systems [21] .They proposed two feature selection algorithms and studied the performance using these algorithms compared to a mutual information based feature selection method.These feature selection algorithms require the use of a featuregood ness measure. Experiments on KDDCup99 dataset address that their proposed mutual information based feature selection method results in detecting intrusions with higher accuracy,especially for remote to login (R2L) and user to remote (U2R) attacks.

III. PROPOSED METHOD

A. CFS and Bayes Theorem

We proposed a new hybrid feature selection method by combining CFS and Bayes Theorem. The CFS algorithm reduces the number of attributes based on the SU measure, In CFS each attributes are compared pair wise to find the Similarity and the Attributes are compared to class attribute to find the amount of contribution it provides to the class value , based on these the attributes are removed. The selected attributes from the CFS algorithm is fed into Bayes theorem for further reduction. Bayes theorem calculates the conditional probability for each attribute and the attribute which has highest conditional probability is selected. Both the Algorithms CFS and Bayes theorem works on the Conditional Probability measure.

B. Proposed Algorithm

```
input: S(F1; F2; :::; FN;C) // a training data set
δ                              // a predefined threshold
output: Sbest {Abest(highest IG)}            // an optimal subset
  begin
     for i = 1 to N do begin
        calculate SUi;c  for Fi;
        if (Su_i,c ≥ δ )
                       append  F_i  to S'_list;
              end;
              order S'_list in descending SUi,c value;
        F_p = getF irstElement(S'_list);
              do begin
       F_q = getNextElement(S'_list , F_p);
       if (F_q <> NULL)
             do begin
                       F'_q = F_q;
              if (SU_p,q , SU_q,c)
                       remove Fq from S'_list ;
              Fq = getNextElement(S'_list , F'_q);
             else Fq = getNextElement(S'_list , F_q);
      end until (Fq == NULL);
  F_p = getNextElement(S'_list , F_p);
 end until (Fp == NULL);
S_best = S'_list ;
Sbest={X1,X2,..XN}
     for j=1 to N begin
          for k=j+1 to N  begin
            P[C_m/(X_j,X_k)]=P[(X_j,X_k)/C_m]*P(C_m)
          P[C/(X_j,X_k)]= P[C_1/(X_j,X_k)]+P[C_2/(X_j,X_k)]+…  ....+P[C_n/(X_j,X_k)]
              If(P[C/Xj,Xk)]>Ω)
                 {
                    if((P[C/Xj]>(P[C/Xk])
                  Remove Xk from Sbest
                  Sbest=IG(X)
                Else
                  Remove Xj from Sbest
                  Sbest=IG(X)

                 }
end;
```

C. Dataset used in the Experiment

The following is the sample of the Heart Disease Data.arff

```
@relation heart-statlog
@attribute age real
@attribute sex real
@attribute chest real
@attribute resting_blood_pressure real
@attribute serum_cholestoral real
@attribute fasting_blood_sugar real
@attribute resting_electrocardiographic_results real
@attribute maximum_heart_rate_achieved real
@attribute exercise_induced_angina real
@attribute oldpeak real
@attribute slope real
@attribute number_of_major_vessels real
@attribute thal real
@attribute class { absent, present}
@data
70,1,4,130,322,0,2,109,0,2.4,2,3,3,present
67,0,3,115,564,0,2,160,0,1.6,2,0,7,absent
57,1,2,124,261,0,0,141,0,0.3,1,0,7,present
64,1,4,128,263,0,0,105,1,0.2,2,1,7,absent
74,0,2,120,269,0,2,121,1,0.2,1,1,3,absent
65,1,4,120,177,0,0,140,0,0.4,1,0,7,absent
56,1,3,130,256,1,2,142,1,0.6,2,1,6,present
59,1,4,110,239,0,2,142,1,1.2,2,1,7,present
```

The Heart Disease data after applying tradional method in Weka,The number high number of attributes reduced is 7 and then these attributes can be fed to various classifiers. The CFS+Bayes theorem algorithm is coded, where the attribute afte**r CFS** is 4 and the selected attributes after Bayes theorem is only 3. CFS Feature selection method which selects the attributes based on the symmetrical uncertainty reduces the number of attributes from 13 to 4.The reduced attributes is fed to Bayes theorem for further reduction.

Table 3.1: Number of original and reduced attributes.

Attribute Selection method	Total Number of Attributes	Number of Attributes after attribute selection
CFS	14	4
CFS+Bayes Theorem	14	3

We applied four classification algorithms for heart disease data such as NB, J48, KNN and NN. First we applied these algorithms for whole data set. The whole data in ARFF document is given to weka and classification algorithm is applied to it.

The heart ARFF will contain large quantity of data and applying classification algorithms to this dataset is time consuming and also gives result with less accuracy. Hence we have to reduce the data set by using attribute selection method. Then this reduced dataset is fed into the four classification algorithm and which algorithm is best fit for this prediction is investigated. Likewise, all other attribute selection and classification algorithms are applied for heart disease dataset. From that we identified that NB classification algorithm gives better accuracy after applying the CFS attribute selection method.

D. Hybrid Feature Selector

The two best Feature Selection methods are applied in sequence. (i.e) CFS followed by Filtered Subset Evaluation. In this method the reduced number of attributes after CFS is 7 and this 7 attributes are fed to Filtered Subset Evaluation which reduces as 6 attributes. After applying the hybrid feature selector, the data is applied to the classification algorithm in which Naïve bayes gives higher Accuracy comparing to the other classifiers.

After applying into Naive bayes, the incorrectly classified instances are separated. The correctly classified samples are kept as training set and the incorrectly classified samples as test set are fed into various other classifiers, where the J48 gives greater accuracy.

E. CFS and Bayes Theorem

We proposed a new hybrid algorithm that is CFS+Bayes Theorem. When applying this feature selection algorithm, the attributes are reduced as 3. Then reduced dataset is given to classifiers. Here Naïve bayes gives the greater accuracy compared to other classifiers.

IV. RESULT AND DISCUSSION

The hybrid feature selector method is automated, where CFS and Bayes theorem's algorithm is coded into program. The number of reduced attributes by CFS and Bayes Theorem is shown in Table 4.1.

Table 4.1: Reduced attributes by CFS and Bayes theorem

Attribute Selection Method	Number of Attributes	Reduced attributes
CFS+Bayes Theorem	14	3

Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data. The Measure of a model's ability to correctly label a previously unseen test case. If the label is categorical (classification), accuracy is commonly reported as the rate which a case will be labeled with the right category. If the label is continuous, accuracy is commonly reported as the average distance between the predicted label and the correct value.

A confusion matrix displays the number of correct and incorrect Predictions made by the model compared with the actual classifications in the test data. The matrix is n-by-n, where n is the number of classes. From that we calculated the accuracy of each classification algorithms.

Table 4.2: Classifiers Accuracy with full dataset.

S.No	Classifiers	Correctly Classified Samples	Incorrectly Classified Samples	Accuracy
1.	Naïve Bayes	226	44	83.70%
2.	J48	207	63	76.66%
3.	KNN	203	67	75.18%
4.	Multilayer	211	59	78.148%

	perception			

A. Attribute Selection

The data set contains the large volume of data. The data are reduced using the attribute selection method. After apply the attribute selection method, the following number of attributes are selected from the whole attributes.

Table 4.3: Number of selected attributes by each attribute selection method.

Attribute Selection Methods	Number of Attributes Selected
CFS subset evaluation	7(3,7,8,9,10,12,13)
Chi-squared attribute evaluation	13(1,2,3,4,5,6,7,8,9,10,11,12,13)
Consistency subset Evaluation	10(1,2,3,7,8,9,10,11,12,13)
Filtered attribute Evaluation	13(1,2,3,4,5,6,7,8,9,10,11,12,13)
Filtered subset evaluation	6(3,8,9,10,12,13)
Gain ratio attribute Evaluation	13(1,2,3,4,5,6,7,8,9,10,11,12,13)
Info gain attribute Evaluation	13(1,2,3,4,5,6,7,8,9,10,11,12,13)
Latent semantic analysis	1(1)
One attribute evaluation	13(1,2,3,4,5,6,7,8,9,10,11,12,13)
Relief attribute evaluation	13(1,2,3,4,5,6,7,8,9,10,11,12,13)

B. Accuracy measure after attribute selection

After reducing the number of attributes, the resulting data is given to the classification and clustering algorithms. It takes less time for computation and improves the accuracy also.

Table 4.4: Accuracy of classifiers with reduced attributes (in %)

Attribute selection Methods	Accuracy of NB	Accuracy of KNN	Accuracy of J48	Accuracy of Multilayer Perceptron	Average
CFS subset eval	85.5	78.14	81.11	82.22	81.74
Chi-squared attribute eval	83.70	75.18	76.66	80.37	78.97
Consistency subset evaluation	84.07	78.14	78.88	81.11	80.55
Filtered attribute Evaluation	83.70	75.18	76.66	80.37	78.97
Filteredsubset eval	85.18	80	79.60	78.88	80.91
Gain ratio attribute Evaluation	83.70	75.18	76.66	78.88	78.60
Info gain attribute Evaluation	83.70	75.18	76.66	80.37	78.97
Latent semantic analysis	54.07	51.11	55.55	52.96	53.42
One attribute eval	83.70	75.18	76.66	79.25	78.69
Relief attribute evaluation	83.70	75.18	76.66	78.14	78.42

1. CFS+Filter Subset Evaluation

Table 4.5: Number of selected features by CFS and Filtered subsetEval

Attribute Selection method	Selected attributes
CFS+FilteredSubsetEval	6(3, 8, 9, 12, 13)

Then this reduced data is given to the classification and clustering algorithms to analyze which is best fit for heart disease prediction.

Table 4.6: Accuracy of classification after CFS+FilteredsubsetEval.

Classification method	Accuracy (in %)
Naïve bayes	85.18
Multilayer perceptron	78.88
KNN	80.74
J48	79.62
Average	83.62

From the above table, we found classification gives the better accuracy. So we take the best classification method on this result.

Table 4.7: Comparison of accuracy of classifiers after CFS+FilteredSubsetEval

Classification Methods	Correctly classified	Accuracy (in %)
NB	230	85.18
Multi layer Perceptron	213	78.88
KNN	218	80.74
J48	215	79.62

2. CFS and Bayes Theorem

We proposed a new hybrid Feature selector combining CFS and Bayes theorem.

Table 4.8: Hybrid feature selection

Attribute Selection methods	Selected attributes
CFS+Bayes theorem	3(3, 12, 13)

Then this reduced data is given to the classification algorithms and calculate the accuracy for identifying the best algorithm.

Table 4.9: Accuracy of classification after CFS+Bayes theorem

Classification algorithms	Accuracy (in %)
Naïve bayes	80.37
Multilayer perceptron	85.18
KNN	85.55
J48	85.18
Average	**84.07**

Table 4.10: Comparison of Accuracy of classifiers with CFS+Bayes theorem as feature selector

Classification method	Correctly classified instances	Accuracy (in %)
NB	**217**	**80.37**
Multi layer perceptron	**230**	**85.18**
KNN	**231**	**85.55**
J48	**230**	**85.18**

Table 4.11: Comparison of feature selection methods

Feature Selection	Accuracy (in %)
CFS	81.74
Filtered Subset eval	81.01
CFS+FilteredSubset eval	83.62
CFS+Bayes Theorem	84.07

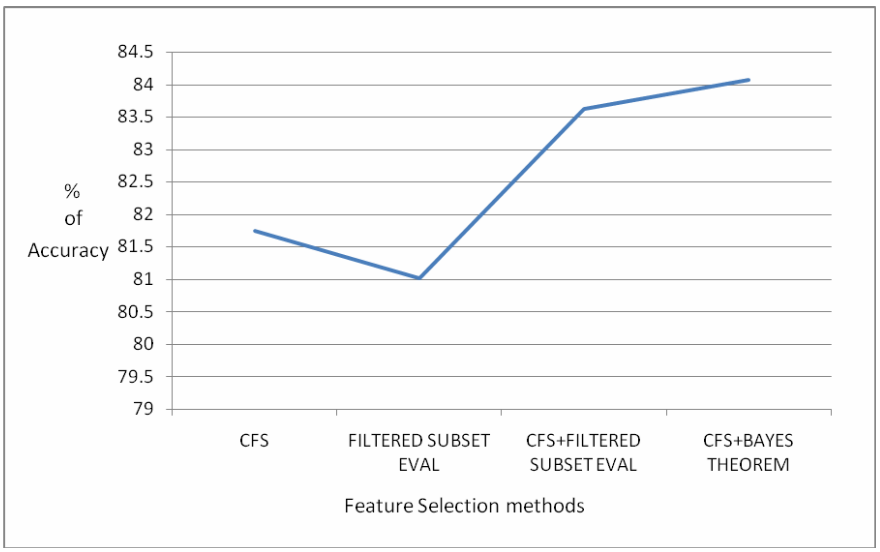

Figure 4.1: Comparison of various feature selectors

From the above investigation, we have to conclude that CFS+Bayes theorem gives the better accuracy compared to the other algorithms.

V. CONCLUSION

This paper investigated the significance of feature selection methods for improving the performance of classification methods. The experimentation is conducted on dataset of health care domain. It is found that the CFS and FILTER SUBSET EVALUATION reduces more number of irrelevant and redundant attributes thereby increases the performance of classifiers. In addition the new feature selection namely CFS and BT was proposed. The proposed algorithm gives better accuracy for NB and KNN classifier. We conclude that **CFS and BAYES THEOREM** based feature selector is best suitable for heart disease data prediction.

REFERENCES

[1] Anamika Gupta, Naveen Kumar, and Vasudha Bhatnagar, "Analysis of Medical Data using Data Mining and Formal Concept Analysis", Proceedings Of World Academy Of Science, Engineering And Technology ,Volume 6, June 2005,.

[2] Andreeva P., M. Dimitrova and A. Gegov, Information Representation in Cardiological Knowledge Based System, SAER'06, pp: 23-25 Sept, 2006.

[3] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

[4] Cristianini, N., Shawe-Taylor, J.: "An introduction to Support Vector Machines. Cambridge University Press", Cambridge, 2000.

[5] Chen, J., Greiner, R.:" Comparing Bayesian Network Classifiers". In Proc. of UAI-99, pp.101–108, 1999.

[6] Compare failure prediction models based on feature selection technique: empirical case from Iran" :Sarah Ashooria, Shahriar Mohammadib. Procedia Computer Science 3 (568–573),2011.

[7] Feature selection for SVM via optimization of kernel polarization with Gaussian,ARD kernels", Tinghua Wanga,b, Houkuan Huang , Shengfeng Tian, Jianfeng Xu. Expert Systems with Applications 37 (6663–6668),2010.

[8] Feature selection for Bayesian network classifiers using the MDL-FS score" Madalina M. Drugan a, Marco A. Wiering . International Journal of Approximate Reasoning 51 (695–717), 2010.

[9] Frank Lemke and Johann-Adolf Mueller, "Medical data analysis using self-organizing data mining technologies," Systems Analysis Modeling Simulation , Vol. 43 , no. 10 , pp: 1399 - 1408, 2003.

[10] Frawley and Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, MenloPark, C.A, 1996.

[11] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[12] Hian Chye Koh and Gerald Tan ,"Data Mining Applications in Healthcare", Journal of healthcare information management, Vol. 19, Issue 2, Pages 64-72, 2005.

[13] Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes , Special report: 2001 – 2003, New Mexico.

[14] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, Knowledge Management, Data Mining, and Text Mining In Medical Informatics", Chapter 1, pgs 3- 34

[15] Importance of feature selection in decision-tree and artificial-neural-networkecological applications Alburnus alburnus alborella: A practical example : Tina Tirelli , Daniela Pessani, Ecological Informatics 6 , 309–315, 2011.

[16] J. Shreve, H. Schneider, O. Soysal:"A methodology for comparing classification methods through the assessment of model stability and validity in variable selection", Decision Support Systems 52 ,247–257,2011.

[17] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences 3; 3, 2008

[18] L. Goodwin, M. VanDyne, S. Lin, S. Talbert ,"Data mining issues and opportunities for building nursing knowledge" Journal of Biomedical Informatics, vol:36, pp: 379-388, 2003.

[19] Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Association Rules. In: Proc. of 2001 International Conference on Data Mining, 2001.

[20] M. Chen, J. Han, and P. S. Yu. Data Mining: An Overview from Database Perspective. IEEE Trans. Knowl. Dat. Eng., vol: 8, no:6, pp: 866-883, 1996.

AUTHORS

First Author – T. John Peter, Dept. of IT, KCG College of Technology, Chennai

Second Author – K. Somasundaram, Dept. of CSE, Jeya College of Engineering, Chennai