A Knowledge driven Approach for Efficient Analysis of Heart Disease Dataset

G. N. Beena Bethel
Associate Professor,
CSE Department,
GRIET, Hyderabad

T. V. Rajinikanth, PhD
Professor,
CSE Department,
SNIST, Hyderabad.

S. Viswanadha Raju, PhD
Professor,
CSE Department,
JNTUH, Jagityal, Karimnagar.

ABSTRACT
Heart Disease Dataset (HDD) contains high dimensions which poses challenges to research community in terms of complexity and efficient analysis. Heart disease is also called as cardiovascular disease (CVD). Feature selection will be made to reduce the irrelevant and redundant number of attributes. Fast diagnosis of the heart disease can be done using a knowledge driven approach. A comparison was made for medically important features to that of computerized subset of features, to bring out much simpler set of features used for the diagnosis. It focuses on the experts' judgement for medical driven feature selection process termed as MFS, and the performance of various classifiers on Cleveland dataset for the computerized feature selection termed as CFS and also a combination of both to enhance the prediction accuracy. Further, this paper categorizes the MFS, CFS and the combination of both into discrete and continuous sets of attributes. Our work has proved that the discrete features do not contribute much to the classification as do the continuous ones, in its accuracy, speed and performance.

Keywords
Medical Feature Selection, Computerized feature selection, SMO

1. INTRODUCTION
Heart disease usually called as Coronary Artery Disease refers to any condition that affects the normal functioning of the heart [1]. CAD is a particular state of heart where the arteries narrow down leading to an obstruction for the blood flow that causes a heart failure which is commonly known as heart attack. Many are the risk factors associated with CAD like sex, age, family history, cholesterol, high blood pressure, diabetes, physical inactivity, obesity, smoking, alcohol consumption, etc., but according to the experts' opinion there are a few important features which are discussed elaborately in section 4 of this paper. J. Nahar et al., in [2] has contributed the medically important features that contribute to the classification of heart disease, and proved that both medically important features as well as computerized features are important for effective classification of heart disease diagnosis. The main objective of our paper is to select the attributes from a combination of both medical and computerized, which will further enhance the prediction accuracy and decision making.

Computation or automation is more likely to eliminate the important features in the process of dimension reduction. But, we still depend on the computerized features for the diagnosis, because of its speed, accuracy, unavailability of Cardiologists in some places, unavailability of diagnostic equipment etc. Therefore this part of the study incorporates some of the medically important features given by expert opinions, so that some justice has been done in bringing out an efficient diagnosing system, which will help the society at large.

The rest of the paper is organized as: Section 2 provides an overview of the existing research work on various computational techniques. Section 3 gives the description of the dataset that has been used and section 4 demonstrates the work that has been done with respect to the proposed methodology. Section 5 illustrates the results obtained in our work and a comparison with the other existing methods. Section 6 draws conclusions for our work and proposed enhancements.

2. EXISTING METHODOLOGY
As per the statistics reported by India Today in 2013, the number one killer disease in India is replaced by cardiovascular diseases, [3]. Varieties of Computational Intelligence techniques have been designed to improve the diagnosis of heart diseases.

Many feature selection methods have been applied in heart disease diagnosis to select the best significant features to diagnose the heart disease. Zhao. H. et al in 2010 [4], have used backward elimination method to identify the biomarkers for unstable angina using metabolites. Wrapper based feature selection using Chi-square statistics was proposed for medical databases by Abraham R. et al in 2007 [5], Sethi P. et al in 2010 [6], made a comparative analysis of Chi-square, gain ratio and information gain on the healthcare data to extract relevant features for classification. Kernel F-score feature selection was used by Polat k. et al in 2009 [7], for the classification of medical databases. A wrapper based feature selection approach was used on conceptual clustering by Mark Devaney et al in 1997 [8].

Other researchers have focused on classification aspects of heart disease diagnosis: Hybrid fuzzy support vector clustering for heart disease identification was proposed by Gamboa A. I. et al in 2006 [9]. Data fusion was used to extract features for heart disease classification and subjected to multi-layer feed forward neural network for classification by Obayya M. et al, in 2008 [10]. Support vector based identification of heart valve diseases using heart sounds was proposed by Maglogiannis et al, in 2009 [11].

The dataset has been split into five distinct subsets, each having one type of class attribute. For each subset of the data, measures like accuracy, True-Positive rate (TP), F-measure and training time were recorded. Accuracy indicates the overall accuracy, TP rate indicates the classification accuracy for positive classes, F-measure indicates the effectiveness of the algorithm, and training time was used to compare the computational complexity of the algorithm.

Table 1. Number of Positive and Negative instances/Subset of records for Healthy, Sick-1, Sick-2, Sick-3 and Sick-4 datasets taken from Cleveland Dataset

Dataset Name	class label considered	Number of positive instances	Number of negative instances
Healthy	0	164	139
Sick-1	1	55	248
Sick-2	2	36	267
Sick-3	3	35	268
Sick-4	4	13	290

K. Usha Rani, [12] gave a description of class label attribute as class 0 as normal person, class1 as first stroke, class 2 as second stroke and class 3 as third stroke and class 4 as end of life. As for our assumption, according to the medical literature, and also from the conclusions drawn from J. Nahar et al., [2], we are able to conform that the attributes required to measure a healthy heart would be Age, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar and Resting ECG. Amma N. G. B. in [13], has categorized the class attributes as Absence (of disease) for 0, Low for 1, Medium for 2, High for 3 and Serious for 4. In our paper, we assume for Sick-1 a low stroke condition, which can be identified with a heart rate variation (Maximum Heart Rate) than the cholesterol level. For Sick-2, we assume a small blockage, which depends on the cholesterol levels. For Sick-3, we assume a medium blockage, which depends both on Maximum Heart Rate and Cholesterol that has high chances of stroke. For Sick-4, we assume a severe blockage which not only depends on maximum heart rate and cholesterol, but also Exercise induced angina to predict the possibility of stroke.

2.1 Computerized Feature Selection (CFS)

The computerized feature selection CFS process was provided by Witten et al., [14] using Weka's CfsSubsetEval attribute selection (uses breadth first strategy). It was observed by Jesmin Nahar et al., [2] that a few medically important features such as age, cholesterol, fasting blood pressure, resting blood pressure and resting ECG were discarded by CFS for 'Healthy' dataset. Similarly, for 'Sick-1' dataset, age, resting blood pressure, fasting blood sugar, maximum heart rate, cholesterol and resting ECG were not considered relevant by CFS. For 'Sick-2' dataset, age, resting ECG, resting blood pressure and cholesterol were considered unimportant by CFS. For 'Sick-3' dataset, resting blood pressure, cholesterol, age, max heart rate and resting ECG were discarded by CFS. For Sick-4 dataset, age, cholesterol, resting blood pressure, max heart rate, resting ECG have been discarded by the CFS. Such outcomes are doubted by medical practitioners and degrades the advantages from computerized system. Table 2 shows the MFS and CFS response over all the five subsets of data.

2.2 Medical Feature Selection (MFS)

A single risk factor taken in isolation cannot figure out all individuals' risk of heart disease. Hence many factors are required to diagnose it. Some factors were suggested in medical literature and some are filtered out using computerized techniques, and Jesmin Nahar et al., [2] has made a study based on the medical literature, and listed out some medically important features. From the following table it is observed that the factors such as cholesterol, heart rate, hypertension (blood pressure), resting ECG, diabetes, blood sugar, stress, exercise induced angina and old age are significant in predicting heart disease. Out of these factors, eight of them are medically significant from the list of Cleveland heart disease dataset. They are age, chest pain type, resting blood pressure, fasting blood sugar, cholesterol, maximum heart rate, resting heart rate and exercise induced angina. If medically significant features were neglected, then it has every chance to run into the risk of incorrect diagnosis.

Considering only Computerised features or only medically important features is insufficient for the diagnosis of heart disease and so both MFS and CFS were taken together.

2.3 MFS + CFS

J. Nahar et al. in paper [2] has experimented and proved that MFS + CFS has resulted in higher performance than the MFS alone or CFS alone. Especially SMO has performed better in terms of accuracy for Healthy, Sick-2, Sick-3 and Sick-4 datasets. Similarly, IBK algorithm also has showed a better performance in terms of TP and F-measure using MFS + CFS. So it is evident from their work that MFS + CFS will give a promising result in the area of classification.

Table 2. Attributes related to MFS, CFS and MFS + CFS for "Healthy", Sick-1, Sick-2, Sick-3 and Sick-4 Datasets respectively as drawn from existing representation

Attributes related to	MFS	CFS	MFS + CFS
Healthy	Age Resting Blood Pressure Cholesterol Fasting Blood Sugar Resting ECG	Old Peak Number of Coloured Vessels Thal	Chest Pain Type Maximum Heart Rate Exercise Ind. Angina
Sick – 1	Age Resting Blood Pressure Maximum Heart Rate Fasting Blood Sugar Resting ECG	Sex Number of Coloured Vessels Thal	Chest Pain Exercise Ind. Angina
Sick – 2	Age Resting Blood Pressure Cholesterol Fasting Blood Sugar Resting ECG	Old Peak Number of Coloured Vessels Thal	Chest Pain Type Maximum Heart Rate Exercise Ind. Angina
Sick – 3	Age Resting Blood Pressure	Slope Number of Coloured Vessels	Chest Pain Type Fasting Blood Sugar

	Cholesterol Maximum Heart Rate Resting EC	Thal	Exercise Ind. Angina
Sick – 4	Age Exercise Ind. Angina Cholesterol Maximum Heart Rate Fasting Blood Sugar	Slope Number of Coloured Vessels Thal	Chest Pain Type Resting ECG

3. DATASET DESCRIPTION

Publicly available heart disease dataset donated by David W. Aha [15], is taken from the UCI repository to work with our experiment (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). Most of the researchers used this Cleveland Dataset for their work and so our work also has adopted the same benchmark dataset. The dataset consists of 76 attributes out of which majority of Computational Techniques have chosen only 14 attributes. The 14 attributes that we have considered along with their details are as follows:

1. Age: in years (continuous);
2. Sex: male or female (discrete);
3. Chest pain type (CP): From medical point of view,
    a. Typical angina (angina), is the condition in which the past history of the patient shows the usual symptoms and so the possibility of having coronary artery blockages is high [16]
    b. Atypical angina (abnang), refers to the condition that the patient's symptoms are not detailed and so the probability of blockages is lower [16].
    c. Non-anginal pain (notang), is the stabbing or knife-like, prolonged, dull, or painful condition that can last for short or long periods of time [16].
    d. Asymptomatic (asympt) pain shows no symptoms of illness or disease and possibly will not cause or exhibit disease symptoms [16].
4. Trestbps: patient's resting blood pressure in mm Hg at the time of admission to the hospital (continuous);
5. Chol: Serum cholesterol in mg/dl; (continuous)
6. Fbs: Boolean measure indicating whether fasting blood sugar is greater than 120 mg/dl: (1 = True; 0 = false) (discrete);
7. Restecg: electrocardiographic results during rest. Three types of values normal (norm), abnormal (abn): having ST-T wave abnormality, ventricular hypertrophy (hyp) (discrete);
8. Thalach: maximum heart rate attained (continuous);
9. Exang: Boolean measure indicating whether exercise induced angina has occurred: 1 = yes, 0 = no (discrete);
10. Oldpeak: ST depression brought about by exercise relative to rest (continuous);
11. Slope: the slope of the ST segment for peak exercise. Three types of values upsloping, flat, downsloping (discrete);
12. Ca: number of major vessels (0–3) colored by fluoroscopy (continuous);
13. Thal: the heart status as retrieved from Thallium test, (normal, fixed defect, reversible defect) (discrete);
14. Num: (class attribute) values are 0 for healthy and 1,2,3,4 for unhealthy.

The Cleveland heart disease dataset has five class attributes indicating either healthy or one of four sick types. For this paper, multi-class classification is converted into a binary classification, thereby it results in 0 for healthy and 1 for unhealthy cases. There are 303 cases in the dataset and has been considered for our work as it was the benchmark dataset. The reason behind using a benchmark dataset is that the comparison of results with other experiments becomes easier and good conclusions can be drawn from such work. Hence Cleveland's heart disease dataset has been used for this experiment.

Table 3. Attributes Related to MFS, CFS, MFS + CFS and MFS + CFS with continuous features for "Healthy", "Sick-1", "Sick-2", "Sick-3" and "Sick-4" Datasets

Attributes related to	MFS	CFS	MFS + CFS	MFS + CFS with Continuous features
Healthy	Age Resting Blood Pressure Cholesterol Fasting Blood Sugar Resting ECG	Old Peak Number of Coloured Vessels Thal	Chest Pain Type Maximum Heart Rate Exercise Ind. Angina	Chest Pain Type Exercise Ind. Angina
Sick – 1	Age Resting Blood Pressure Maximum Heart Rate Fasting Blood Sugar Resting ECG	Sex Number of Coloured Vessels Thal	Chest Pain Exercise Ind. Angina	Chest Pain Type Exercise Ind. Angina
Sick – 2	Age Resting Blood Pressure Cholesterol Fasting Blood Sugar	Old Peak Number of Coloured Vessels Thal	Chest Pain Type Maximum Heart Rate Exercise Ind. Angina	Chest Pain Type Exercise Ind. Angina

	Resting ECG			
Sick – 3	Age Resting Blood Pressure Cholesterol Maximum Heart Rate Resti ECG	Slope Number of Coloured Vessels Thal	Chest Pain Type Fasting Blood Sugar Exercise Ind. Angina	Chest Pain Type Fasting Blood Sugar Exercise Ind. Angina
Sick – 4	Age Exercise Ind. Angina Cholesterol Maximum Heart Rate Fasting Blood Sugar	Slope Number of Coloured Vessels Thal	Chest Pain Type Resting ECG	Chest Pain Type Resting ECG

4. CONTINUOUS ATTRIBUTES

There are basically two types of attributes in datamining like discrete and continuous. Discrete attributes are those which have finite or countable set of values. Examples of discrete data are pin-codes, sex of a person, numbers on a dice, etc. Continuous data on the other hand has real numbers as values of attributes, widely represented as floating point numbers. An emphasis is given to continuous attributes as they are practically measurable and can be represented with finite number of digits. Examples of continuous data includes temperature, height, length, age of a person etc.

The data that is fixed with respect to time is what we called discrete and such data's contribution towards the decision making is far less than that of the continuous data where even a very few such attributes contribute more towards the decision making. For example, out of thirteen attributes listed above in Cleveland's heart disease data, the attributes drawn from MFS + CFS were Chest Pain Type, Exercise Induced Angina and Maximum Heart Rate, out of which Maximum heart rate is a discrete attribute derived from the Exercise Induced Angina, and nothing much to contribute for the work. So the emphasis has been given to the continuous features.

Cleveland heart disease data has got both continuous and discrete types of data. As on date many researchers are working on discretization of continuous data. Lukasz A. Kurgan and Krzysztof J. Cios [17], members of IEEE, have worked on discretization based on CAIM (class-attribute interdependence maximization), which is designed to work with supervised data. It works on maximizing the class-attribute interdependence to generate a (possibly) minimal number of discrete intervals. Usama M. Fayyad and Keki B Irani [18], have worked on entropy minimization heuristic based on minimum description length principle for discretizing the continuous valued attribute. Richard Butterworth, Dan A. Simovici, Gustavo S. Santos and Lucila Ohno-Machado [19], have worked on a greedy algorithm for supervised discretization of continuous values. James Dougherty, Ron Kohavi, Mehran sehami [20], have worked on supervised and unsupervised discretization of continuous features assuming a Gaussian distribution. Cheng-Jung Tsai, Chien-I. Lee, Wei-Pang Yang [21] have contributed Class-Attribute Contingency Coefficient (CACC) algorithm for discretization of continuous values. Ying Yang and Geoffrey I. Webb [22], have focused on developing a Proportional k-interval discretization algorithm. Discretization algorithms on the whole was proposed by H. Liu, F. Hussain, C.L. Tan, M. Dash, [23] as five different axes which can be classified as supervised versus unsupervised, static versus dynamic, global versus local, top-down (splitting) versus bottom-up (merging), and direct versus incremental. But the research using continuous features is very negligible or almost NIL. Further, David Kashmer, chair of Surgery at Signature Healthcare suggests his healthcare colleagues to use continuous data, as it improves the quality of the project in doing lot more with lot less of it.

Our assumption in this work was if discrete features (like sex of the patient) remains same for any period of time, then such attributes do not contribute much in the prediction. For this cause, we tried testing the accuracy of MFS + CFS over continuous features leaving the discrete features. The features like age for example are considered, the accuracy of prediction varies along with time. As the time passes, the age increases and the risk associated with the heart disease is expected to increase. Along with age, factors like blood pressure at rest, fasting blood sugar, cholesterol, exercise induced angina may also vary with respect to age. Thus, if continuous attributes are used, the robustness of the prediction may increase. The accuracy, sensitivity, specificity, recall, precision and F-measure have shown up far better with MFS + CFS having only continuous features compared to taking all the attributes. We therefore conclude that the discrete features have a very less contribution to the prediction than that of continuous attributes and so we extended the existing data representation to a little more than considering only medically selected features.

SVM classifier has proved to be an efficient classifier in most of the research and we therefore used the Sequential Minimal Optimization method of the SVM classifier to build the classification model. The coding part was done using Matlab as it gives way for a better accuracy. Bar charts were plotted for each of the measures in all the five sub-datasets and only "Healthy" observations were listed out in Fig. 1

5. RESULTS AND COMPARISONS

The accuracy of the feature selection method will be enhanced if there is a control over the data through discrete features. In this work, the accuracy, sensitivity, specificity, recall, precision and F-measure for MFS, CFS, MFS+CFS and MFS+CFS with continuous features taking all the sub-datasets, "Healthy", "Sick-1", "Sick-2", "Sick-3" and "Sick-4" were calculated and the values that are achieved through this work are shown in table 4. It is observed from this table that either CFS or MFS+CFS or both have shown up an improved accuracy when only the continuous features were considered in all the sub-datasets. The features like Chest Pain Type and Exercise Induced Angina were contributing more to classify "Healthy", "Sick-1" and "Sick-2" sub-datasets, Chest Pain Type, Exercise Induced Angina and Fasting Blood Sugar proved to contribute more towards classifying "Sick-3" sub-dataset and Chest Pain Type and Resting ECG contributed more towards classifying "Sick-4" sub-dataset. This shows that continuous attributes contribute more than the discrete features for the classification task. In the above figure, Accuracy, Sensitivity, Specificity, Recall, Precision, F-measure have greatly improved, and the Processing time and Overhead have greatly reduced for continuous features over MFS+CFS.

Table 4. Measurement of Accuracy, Specificity, sensitivity, Recall, Precision, F-measure Time and Overhead with respect to MFS, CFS and MFS+CFS of all the attributes versus only continuous attributes.

	Feature Selection Method	Accuracy	Sensitivity	Specificity	Recall	Precision	F-Measure	Time	Overhead
	MFS	55.670103	0.6	0.510638	0.6	0.566038	0.582524	0.745568	200
	CFS	54.639175	0.82	0.255319	0.82	0.539474	0.650794	0.148939	120
Healthy	MFS + CFS	53.608247	0.76	0.297872	0.76	0.535211	0.628099	0.138992	120
	MFS	49.484536	0.58	0.404255	0.58	0.508772	0.542056	0.273163	**120**
	CFS	**71.134021**	**0.88**	**0.531915**	**0.88**	**0.666667**	**0.758621**	**0.143835**	**40**
Healthy Continuous	MFS + CFS	50.515464	0.7	0.297872	0.7	0.514706	0.59322	0.13704	**40**
	MFS	55.670103	0.7	0.404255	0.7	0.555556	0.619469	0.310743	200
	CFS	60.824742	0.84	0.361702	0.84	0.583333	0.688525	0.139361	120
Sick-1	MFS + CFS	48.453608	0.72	0.234043	0.72	0.5	0.590164	0.132704	80
	MFS	48.453608	0.7	0.255319	0.7	0.5	0.583333	0.374383	**120**
	CFS	39.175258	0.56	0.212766	0.56	0.430769	0.486957	0.183035	**40**
Sick-1 Continuous	MFS + CFS	**68.041237**	**0.9**	**0.446809**	**0.9**	**0.633803**	**0.743802**	**0.132998**	**40**
	MFS	55.670103	0.6	0.510638	0.6	0.566038	0.582524	0.429107	200
	CFS	54.639175	0.82	0.255319	0.82	0.539474	0.650794	0.1429	120
Sick-2	MFS + CFS	53.608247	0.76	0.297872	0.76	0.535211	0.628099	0.152253	120
	MFS	49.484536	0.58	0.404255	0.58	0.508772	0.542056	0.295342	**120**
	CFS	**71.134021**	**0.88**	**0.531915**	**0.88**	**0.666667**	**0.758621**	**0.173938**	**40**
Sick-2 Continuous	MFS + CFS	**50.515464**	**0.7**	**0.297872**	**0.7**	**0.514706**	**0.59322**	**0.159521**	**40**
	MFS	55.670103	0.7	0.404255	0.7	0.555556	0.619469	0.358246	200
	CFS	54.639175	0.82	0.255319	0.82	0.539474	0.650794	0.15915	120
Sick-3	MFS + CFS	48.453608	0.72	0.234043	0.72	0.5	0.590164	0.147288	80
	MFS	48.453608	0.7	0.255319	0.7	0.5	0.583333	**0.365719**	**120**
	CFS	**71.134021**	**0.88**	**0.531915**	**0.88**	**0.666667**	**0.758621**	**0.160317**	**40**
Sick-3 Continuous	MFS + CFS	**68.041237**	**0.9**	**0.446809**	**0.9**	**0.633803**	**0.743802**	0.138396	**40**
	MFS	61.85567	0.82	0.404255	0.82	0.594203	0.689076	0.684797	200
	CFS	54.639175	0.82	0.255319	0.82	0.539474	0.650794	0.14792	120
Sick -4	MFS + CFS	48.453608	0.72	0.234043	0.72	0.5	0.590164	0.13119	80
	MFS	48.453608	0.7	0.255319	0.7	0.5	0.583333	0.316849	**120**
	CFS	**71.134021**	**0.88**	**0.531915**	**0.88**	**0.666667**	**0.758621**	0.146043	**40**
Sick-4 Continuous	MFS + CFS	**68.041237**	**0.9**	**0.446809**	**0.9**	**0.633803**	**0.743802**	**0.146812**	**40**

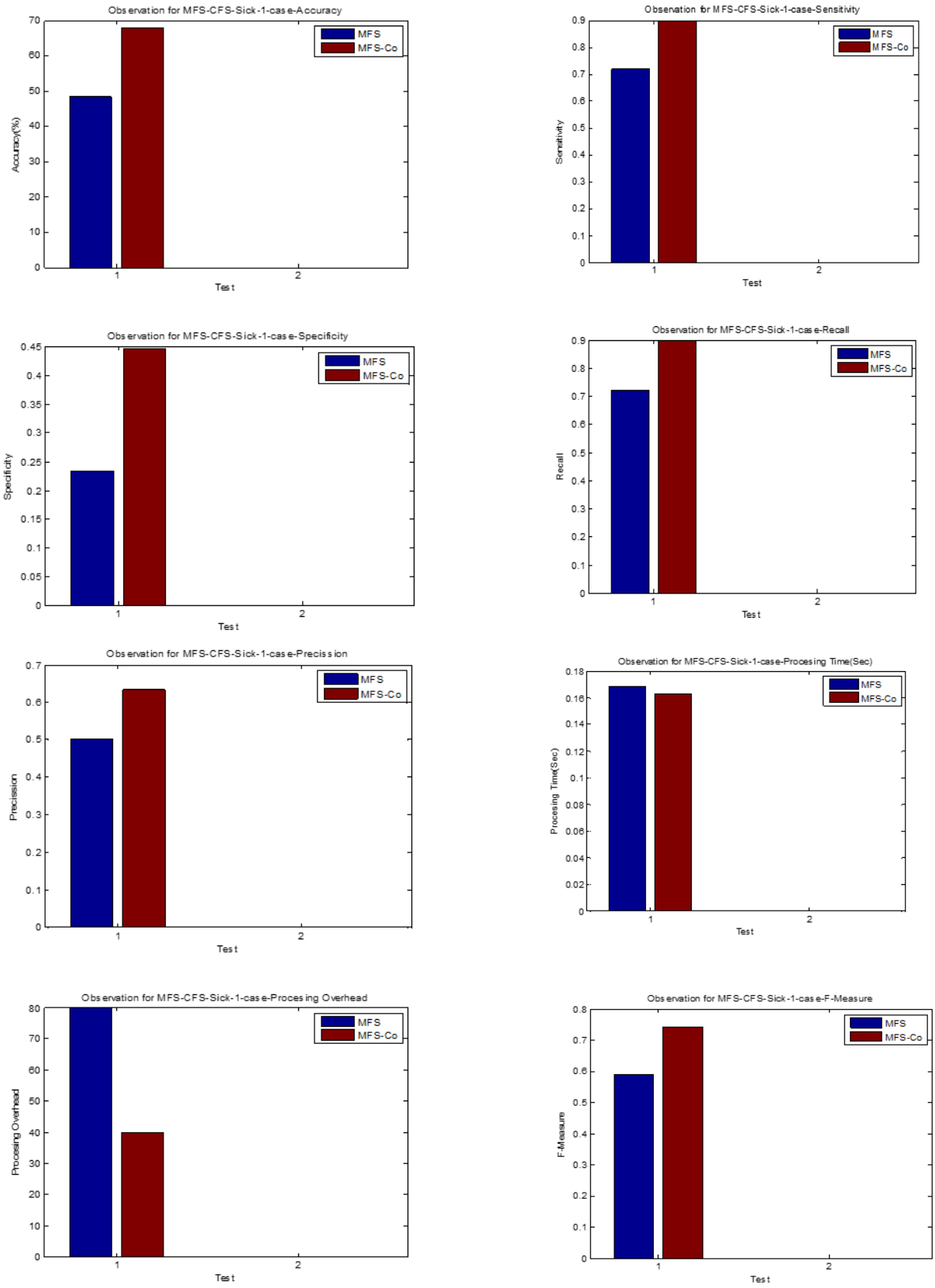

Fig. 1 Analysis in the form of bar charts are shown for one type of dataset "Sick-1" considering the measures as Accuracy, Sensitivity, Specificity, Recall, Precision, F-Measure, Time taken and Overhead are plotted using continuous features of MFS+CFS

6. CONCLUSIONS

Data controlling by minimizing the number of discrete features results in high accuracy and a minimum overhead. From the above analysis it is quite clear that the accuracy, sensitivity, specificity, recall and precision have increased whereas processing time and overhead have decreased enormously when the CFS or MFS+CFS or both have considered **only** continuous features for classification. Almost for all the combinations, reducing the discrete features greatly decreases the processing time and overhead. Therefore we can conclude that by controlling some of the discrete features, the accuracy of the classification is improved and that the continuous features contribute more towards the efficiency of the classification.

7. REFERENCES

[1] http://yourtotalhealth.ivallage.com/heart-disease-fast-facts.html, 2008.

[2] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen (2013), "Computational Intelligence for heart disease diagnosis: A medical knowledge driven approach", Expert system with Applications, 40, 96-104.

[3] www.indiatoday.intoday.in › India, a report on statistics of causes of death in India 2013.

[4] Zhao, H., Guo, S., Chen, J., Shi, Q., Wang, J., Zheng, C., et al. (2010). Characteristic pattern study of coronary heart disease with blood stasis syndrome based on decision tree. In 4th international conference on bioinformatics and biomedical engineering (iCBBE) (pp. 1–3). Chengdu, China: IEEE.

[5] Abraham, R., Simha, J. B., & Iyengar, S. (2007). Medical datamining with a new algorithm for feature selection and Naı¨ve Bayesian classifier. In 10th international conference on information technology, (ICIT), 2007 Orissa IEEE computer society (pp. 44–49).

[6] Sethi, P., & Jain, M. (2010). A comparative feature selection approach for the prediction of healthcare coverage. Information Systems, Technology and Management, 392–403.

[7] Polat, K., & Guenes, S. (2009). A new feature selection method on classification of medical datasets: Kernel F-score feature selection. Expert Systems with Applications, 36, 10367–10373.

[8] Devaney, M., & Ram, A. (1997). Efficient feature selection in conceptual clustering. In Proceedings of the fourteenth international conference on machine learning, Nashville, TN, Citeseer (pp. 92–97).

[9] Gamboa, A. L. G., Mendoza, M. G., Orozco, R. E. I., VARGAS, J. M., & Gress, N. H. (2006). Hybrid Fuzzy-SV clustering for heart disease identification, computational intelligence for modelling. In International conference on control and automation, 2006 and international conference on intelligent agents, web technologies and internet commerce (pp. 121–121).

[10] Obayya, M., & Abou-chadi, F. (2008). Data fusion for heart diseases classification using multi-layer feed forward neural network. In International conference on computer engineering & systems, ICCES (Vol. 978, pp. 6–70).

[11] Maglogiannis, I., Loukis, E., Zafiropoulos, E., & Stasis, S. (2009). Support vectors machine-based identification of heart valve diseases using heart sounds. Computer Methods and Programs in Biomedicine, 95, 47–61.

[12] K. Usha Rani (2011), "Analysis of Heart Diseases Dataset using Neural Network Approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5.

[13] Amma N G B(2012), "Cardiovascular Disease Prediction System using Genetic Algorithm and Neural Network", International Conference on Computing, Communication and Applications (ICCCA), IEEE Explore.

[14] Witten, I. H., & Frank, E. (2005). "Data mining: Practical machine learning tools and techniques." San Francisco: Morgan Kaufmann.

[15] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."

[16] Baliga, R. R., & Eagle, K. A. (2010). "Practical cardiology: Evaluation and treatment of common cardiovascular". Lippincott Williams & Wilkins.

[17] Lukasz A. Kurgan and Krzysztof J. Cios, Members of IEEE, (2004), "CAIM Discretization Algorithm", IEEECS Log Number 114171.

[18] U. M. Fayyad, K. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proc. of the 12th International Joint Conference on Artificial Intelligence, 1993, pp. 1022-1027.

[19] Richard Butterworth, Dan A. Simovici, Gustavo S. Santos and Lucila Ohno-Machado, 2004, "A greedy algorithm for supervised discretization", Elsevier Science.

[20] James Dougherty, Ron Kohavi, Mehran sehami, "Supervised and unsupervised discretization of continuous features".

[21] Cheng-Jung Tsai, Chien-I. Lee, Wei-Pang Yang (2008), "A Discretization algorithm based on Class-Attribute Contingency Coefficient", Information Sciences, 714–731, Elsevier.

[22] Ying Yang and Geoffrey I. Webb, "Proportional k-interval discretization for Naive-Bayes Classifiers", in proceedings of 12 European Conference on Machine Learning (ECML01), pp 564-575.

[23] H. Liu, F. Hussain, C.L. Tan, M. Dash, (2002), "Discretization: an enabling technique", Journal of Data Mining and Knowledge Discovery, 393–423.

[24] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufman Publishers, 2009.

[25] Takeharu Hayashi, MD, PHD, Takuro Arimura, DVM, PHD, Manatsu Itoh-Satoh, MD, PHD, Kazuo Ueda, MD, Shigeru Hohda, MD, PHD, Natsuko Inagaki, MD, Megumi Takahashi, MS, Hisae Hori, PHD, Michio Yasunami, MD, PHD, Hirofumi Nishi, MD, PHD, Yoshinori Koga, MD, PHD, Hiroshi Nakamura, MD, PHD, Masunori Matsuzaki, MD, PHD, Bo Yoon Choi, MS, Sung Won Bae, PHD, Cheol Woon You, MD, Kyung Hoon Han, MD, Jeong Euy Park, MD, Ralph

Knöll, MD, PHD, Masahiko Hoshijima, MD, PHD, Kenneth R. Chien, MD, PHD, Akinori Kimura, MD, PHD, "Tcap Gene Mutations in Hypertrophic Cardiomyopathy and Dilated Cardiomyopathy", Journal of the American College of Cardiology, Vol. 44, No. 11, 2004.

[26] Tommy Jönsson, Yvonne Granfeldt, Bo Ahrén, Ulla-Carin Branell, Gunvor Pålsson, Anita Hansson, Margareta Söderström and Staffan Lindeberg, "Beneficial effects of a Paleolithic diet on cardiovascular risk factors in type 2 diabetes: a randomized cross-over pilot study", Cardiovascular Diabetology, 2009.

[27] Valentin Fuster, MD, PHD, FACC, Pedro R. Moreno, MD, FACC, Zahi A. Fayad, PHD, FACC, Roberto Corti, MD, FACC, Juan J. Badimon, PHD, FACC, "Atherothrombosis and High-Risk Plaque Part I: Evolving Concepts", Vol. 46, No. 6, Journal of the American College of Cardiology, 2005.

[28] Dariush Mozaffarian, MD, Dr PH; Peter W.F. Wilson, MD; William B. Kannel, MD, MPH, "Beyond Established and Novel Risk Factors Lifestyle Risk Factors for Cardiovascular Disease"**,** http://circ.ahajournals.org, American Heart Association, 2008.