An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease

Vinitha DOMINIC [1], Deepa GUPTA [2,*], and Sangita KHARE [1]

[1] Department of Computer Science, Amrita School of Engineering Bangalore Campus, Amrita Vishwa VidyaPeetham, Kasavanahalli, Carmelaram P.O., Bengaluru-560035, Karnataka, India.
[2] Department of Computer Mathematics, Amrita School of Engineering Bangalore Campus, Amrita Vishwa VidyaPeetham, Kasavanahalli, Carmelaram P.O., Bengaluru-560035,Karnataka, India.
E-mails: vinnismail@gmail.com; g_deepa@blr.amrita.edu; k_sangita@blr.amrita.edu

* Author to whom correspondence should be addressed; Tel.: 9916921850

Received: 16 March 2015/Accepted: 28 March 2015/Published online: 31 March 2015

Abstract
Machine learning techniques will help in deriving hidden knowledge from clinical data which can be of great benefit for society, such as reduce the number of clinical trials required for precise diagnosis of a disease of a person etc. Various areas of study are available in healthcare domain like cancer, diabetes, drugs etc. This paper focuses on heart disease dataset and how machine learning techniques can help in understanding the level of risk associated with heart diseases. Initially, data is preprocessed then analysis is done in two stages, in first stage feature selection techniques are applied on 13 commonly used attributes and in second stage feature selection techniques are applied on 75 attributes which are related to anatomic structure of the heart like blood vessels of the heart, arteries etc. Finally, validation of the reduced set of features using an exhaustive list of classifiers is done.In parallel study of the anatomy of the heart is done using the identified features and the characteristics of each class is understood. It is observed that these reduced set of features are anatomically relevant. Thus, it can be concluded that, applying machine learning techniques on clinical data is beneficial and necessary.

Keyword: Data Mining; Genetic Search (GS); InfoGain (IG); Classifier; Cardiovascular Disease

Introduction

Data mining is a new approach to data analysis and information discovery. Data mining is the analysis of large datasets to find hidden relationships and discover useful pieces of information [1], by applying machine learning techniques. Information technology has evolved over the recent years, this is evident in all domains. One among them is the healthcare domain. There is a growing trend of applying machine learning techniques on medical data. The healthcare domain generates exchanges and stores a multitude of patient-specific data. Information technology has transformed the way this data is stored and documented. One of the significant evolutions is that of storage of healthcare data in standardized format structure i.e. an electronic health record which has helped in better research on this data [2]. Data mining techniques can be applied to healthcare domain in order to catalyze and support goals like bypassing clinical trails, finding adverse drug reactions, reducing hospital acquired infections, and rooting out fraud. Data mining algorithms have three categories classification, association and clustering. Each of these categories has different guidelines

Copyright ©2015 by the authors; Licensee SRIMA, Cluj-Napoca, Romania.

on how they can be applied on clinical data [3]. Efficient and systematic use of these data mining techniques will help discover significant information.

The clinical data has various domains like cardiovascular diseases, cancer, diabetes and drugs. Each domain has various attributes and different types of data, where data can be numerical or categorical. Studying each domain requires different guidelines to be followed based on its relevance and significance. In the recent years, there has been a growing impact of cardiovascular diseases on society, leading to high global mortality rate. There is a growing trend of these diseases in low and middle income countries [4]. Early prevention and diagnosis of these diseases could improve the quality of life. This was one of the significant reasons why this paper focused on applying machine learning techniques on cardiovascular diseases dataset. There were two kinds of research workloads reported for cardiovascular diseases, i.e. a classification/prediction model and the other was based on dimension reduction to improve accuracy. In first category of research workload, it was observed that various prediction models were proposed, which used 13 features and binary class i.e. 0 for absence and 1 for presence of heart disease. These prediction models were built using the data mining techniques like Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and Association rules [5-7]. Some workload was also done using Computational intelligence techniques like Neural networks and Multilayer Perceptron for predicting cardiovascular disease. In all the above mentioned research workloads Cleveland dataset [8] provided by the UCI repository is mostly used.. The UCI machine learning repository provided three other datasets namely Hungarian dataset [9], LongBeach dataset [10] and Switzerland dataset [11]. Research had been carried out on these datasets, i.e. certain modified approaches to computational intelligence methods like neural networks and multi layer perceptron for prediction of heart diseases were proposed and tested [12-14]. Second category of research workload done on this dataset was reducing the 13 attributes like age, sex, blood sugar etc. Efforts were made to apply feature selection techniques and improve the accuracy of the classifiers. GS was one of the feature selection techniques that was used and a reduction of features from 13 to 6 was observed [15]. Furthermore, effort had been made to reduce these 6 attributes to 4 attributes using fuzzy rules which are a set of distinct elements with different degree of relevance or membership and takes values between 0 and 1. These were then tested on locally available dataset, with significant improvement in accuracy [16]. Some research gaps found in the past work was that the work had been carried out only on the binary class and on 13 attributes out of 75 available attributes. Though the class attribute had values 0 for absence and 1,2,3,4 for the various risk levels of heart disease, only binary class had been used. Although 75 attributes had been available, feature reduction techniques were used only to reduce these 13 attributes without any valid reason. In spite of availability of exhaustive list of classifiers, only few like NB, DT, SVM and Association rules or combination of these were used for validation.

From the literature survey conducted, it is found that in the past workloads an accuracy of more than 95% was observed because classifiers work well with binary class. It would be more beneficial if the diseases were predicted based on the level of risk rather than its presence or absence. After a clear understanding of these research gaps, the main focus in this research was to study all the four available heart disease datasets with 75 attributes and five classes {0,1,2,3,4}, to apply two feature selection techniques i.e. IG and GS on these features to get a reduced set of features. Finally, validation of these reduced set of features on a exhaustive set of classifiers in terms of accuracy and anatomic relevance was conducted.

Methodology

Proposed Approach

The proposed approach is as follows, the dataset was preprocessed i.e. missing data was removed and the dataset was converted to *arff* format (Attribute-Relation File Format). The next step was to apply IG and GS feature selection techniques on this dataset and a reduced set of features were obtained. Each feature selection technique would give a unique reduced set of features. Each of these unique feature set was tested for accuracy with 6 different classifiers i.e. NB, DT, SVM, Logistic Regression (LR), Multi Layer Perceptron (MLP) and Adaboost. Similar validation was done for the common set of attributes obtained from both feature selection technique across all the datasets. The last step was validation and analysis of the reduced features based on their anatomic relevance. A diagrammatic representation of this approach is shown in Figure 1.

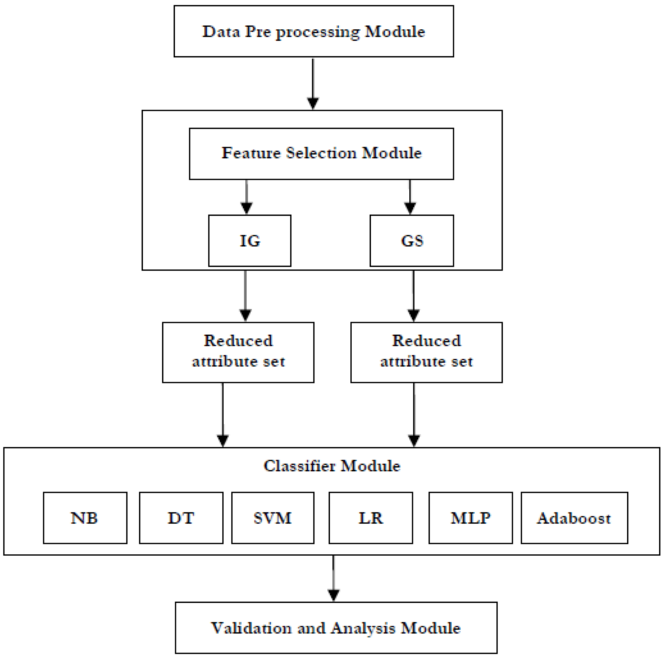


Figure 1. Schematic Diagram of Proposed Approach

Feature Selection Techniques

Two feature selection techniques GS and IG were used. GS introduced the principle of evolution and genetics, among possible solutions to the given problem. Genetic Search algorithm tries to mimic human behavior. The GS approach starts by considering the initial sample of attributes as chromosomes and the best among these chromosomes (attributes) are selected for feature selection. A fitness function was derived for evaluation of these individuals. This was then followed by the process of crossover and mutation. The above process was repeated for determined number of generations, and results were analyzed [17]. IG evaluated the worth of an attribute by measuring the information gain with respect to the class [18]. It used the ranker method which was used to filter the attributes and rank them for selection. Through repeated iterations of changing the ranking parameters, the reduced attributes set could be obtained.

Classifier

Based on the popularity and performance, an exhaustive list of classifiers was chosen for the validation of the reduced set of features. The classifiers selected for this work were NB, DT uses J48 algorithm, SVM, LR, MLP and Adaboost. These classifiers were found to work mostly with nominal values. Naïve Bayes classifier is one of the most popular and frequently used techniques in data mining. It is based on conditional probability of the features and can handle multiple classes. It helps to understand the conditional independence of the attributes, which is one of the reasons for selecting this classifier [19]. DT is another popular technique which helps in understanding the hierarchical significance of each attribute. Using this classifier the attribute which was most useful to identify the class could be obtained and dip in its performance indicates that there was a possibility of over fitting in data. The next classifier used in this paper was SVM called the low error classifier. It gives mathematical validation for classification and uses a margin i.e. a mathematically defined boundary to classify. It was found to work well for binary classes.

It was dependent on the kernel function, thus it eases to identify the best kernel function for the datasets. LR is a significant classifier which uses an equation for all the features and it can predict the class. It focuses on finding the best fit parameters and thus train the classifier. This classifier is highly sensitive to under fitting of data which leads to low accuracy. The next classifier Multilayer Perceptron is based on computational intelligence techniques. It uses back propagation technique for classification of instances. It has various parameters which can be varied for each iteration. It is a popular technique which yields good accuracy and effective validation of the selected attributes. The last classifier in the list is Adaboost which is a meta-algorithm i.e. combining other algorithms. It is considered as the best-supervised learning algorithm, which can work with almost all classifiers, has low error rate and easy to implement with no varying parameters. It was observed to have better accuracy compared to other classifiers, which made it the most significant classifier for analysis and using this classifier for validation could not be neglected [20]. A major dip in performance was noted if there were outliers in the datasets.

Data Statistics

As mentioned earlier there were four datasets investigated, i.e. Cleveland dataset, Hungarian dataset, LongBeach dataset and Switzerland dataset. The datasets had been preprocessed before they were used i.e. missing data were removed. The UCI had claimed that each of the datasets had 303 instances and 75 attributes which included the class attribute which had the following values 0, 1, 2, 3, 4 where class 0 was for absence and 1, 2, 3, 4 for the various risk levels of cardiovascular disease. After the preprocessing step the number of instances available for each of the datasets is shown in Table1. It also shows the data distribution of each class for each dataset. It was observed that the Cleveland and Hungarian datasets have uniform distribution of the instances of healthy individuals and those having cardiovascular disease. But the LongBeach and Switzerland datasets have skewed distribution of instances. Table 1 shows that in the Cleveland dataset majority of the instances (56%) are healthy individuals (class 0) and remaining are unhealthy individuals distributed among the four classes.

Table 1. Data statistics for each dataset

Dataset	Total # of instances	C0	C1	C2	C3	C4
Cleveland	283	157	50	31	32	11
Hungarian	294	188	37	26	28	15
LongBeach	200	51	56	41	42	10
Switzerland	123	8	48	32	30	5

Similar distribution is observed in the Hungarian dataset where 64% patients are healthy. The data distribution for the LongBeach and Switzerland datasets is slightly different, where majority of the instances are unhealthy. In the Switzerland dataset approximately 93% of instances are unhealthy. Another observation is that in all the datasets class 2 and 3 have similar distribution of the number of instances and class 4 has the least number of instances. Preprocessed data is given to the Weka tool for analysis, the description of which is given in the following section.

WEKA Tool

For the implementation of the proposed approach, a Software Tool called Weka was used [21]. It is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. WEKA provides a number of feature selection techniques; this paper uses two of them i.e. IG and GS. In case of feature selection techniques the main parameter that determines the selection is search method, values for these non variable parameters are set as shown in Table 2.

The attribute evaluator for IG is ranker, this method has a parameter called rank where the upper bound of rank of the attributes required can be specified and can be varied for various iterations. Similarly GS has some variable parameters that can be varied. Table 2 shows default parameter settings. When feature reduction was done, each dataset had their own settings. The Cleveland and LongBeach datasets had a crossover rate of 0.8 while it is 0.6 in the case of the Hungarian and Switzerland datasets. All the four datasets have the same mutation rate of 0.033. Number of generations for the Hungarian and Switzerland datasets are 200 and for the Cleveland and LongBeach datasets is 150. Similarly, Table 3 shows the parameter setting for each classifier. In Weka every classifier has different parameters based on its behaviour, for instance SVM has a parameter to set the kernel function which can take Rbfkerne(radial basis function kernel) Linearkernel(linear kernel), Polykernel(polynomial kernel) etc. For this analysis Polykernel was found to perform well and thus the parameter kernel is set to PolyKernel. Similarly other classifiers also have some parameters with default values and variable values. It was observed NB and LR were set to default values. DT had variable values for parameters related to pruning and MLP had for the parameter learning rate. Adaboost had parameter to set the classifier, this parameter had been set for all the six classifiers in the proposed approach and only the accuracy of the best classifier is reported.

Table 2. Parameter for feature selection technique

Feature selection technique	Basic Parameters
IG	Search method = Ranker; Binarize numeric attributes = False; Missing merge = True
GS	Attribute evaluator = CFS Subset Eval; Crossoverrate = 0.6; Mutationrate = 0.033; Num of generations = 20; Population size = 20; Seed=1

Table 3: Parameters for classifers in WEKA

Classifier	Basic Parameters
NB	Use Kernel Estimator = False; Use Supervised Discretization = False
DT	Binary Splits = False; Confidence factor = 0.25; Min Num Obj = 2; Num Folds = 3; Reduced Error Pruning = False; Save Instance Data = False; Seed = 1; Subtree Raising = True; Unpruned = False; Use Laplace = False
SVM	Build Logistic Models = False; c=1.0; Checks Turned Off = False; Epsilon =1.0E-12; Kernel = Polykernel; Num of Folds = 1; Randomnseed = 1; Tolerance parameter = 0. 001
LR	Max Its = 1; Ridge = 1.0E-8
MLP	Hidden layers = a; Learning Rate = 0.3; Momentum =0.2; Nominal To Binary Filter = True; Normalize Attributes = True; normalizeNumericClass=True; Reset = True; Seed = 0; Trainingtime = 300; Validation Threshold = 20
Ada-Boost	Classifier = J48; Num Iterations = 10; Seed = 1; Use Resampling = False; Weight Threshold = 100

The usual procedure for diagnosis of cardiovascular disease by a medical practitioner involves three stages. First is the test for risk factors i.e. blood cholesterol levels, diabetes, blood pressure, patient demographics like age, smoking habits etc. This is followed by stress testing, an electrocardiogram and finally a coronary angiography [22]. The 75 attributes in a cardiovascular disease dataset are related to these three diagnosis procedures. The description of these attributes is available in the UCI repository. The main focus of the analysis is to understand how these attributes help in diagnosis. The 13 commonly used attributes with binary class help to understand the presence or absence of cardiovascular disease, but the interest of an individual is to understand the risk level, based on this the corrective measures and medication can be suggested. Keeping this goal in mind, the analysis is carried out in two stages, first using the 13 commonly used attributes and 5 classes, then using 75 attributes and 5 classes

Experimental Results

In the first stage of analysis the 13 attributes such as age, sex, cholesterol (*chol*), fasting blood sugar (*fbs*), exercise test values (*thal*, *thalch*, *thaltime*, chest anigma (*ca*), resting blood pressure (*trestthps*)). Resting electrocardiographic results (*restecg*), etc. and classes C0-C4 were analyzed. The UCI repository gives a detailed description of these attributes. The feature selection techniques IG and GS were applied, list of the reduced set of attributes is shown in Table4**.** It can be observed that the attribute *chest pain (cp)* is common across all datasets after applying IG. Similarly for GS *exang (exercise anigma)* is found common. It is also observed that both the LongBeach and Switzerland datasets have the same attribute *exang* after applying GS. So thus the feature *exang* i.e. the variation in the pulse observed during exercise is a prominent feature among the 13 features. Further, these reduced features were validated using 6 classifiers depicted in Figure 2.

Table 4. Selected attributes from 13 attributes using IG and GS

Feature Selection Technique	# of selected attributeses	List of attributes
Clevland Dataset		
IG	5	thal, cp, ca, exang, oldpeak
GS	8	sex, cp, thalach, exang, oldpeak, slope, ca, thal
Hungarian Dataset		
IG	8	slope, exang, cp, oldpeak, thalach, sex, trestbps, thal
GS	4	sex, cp, exang, slope
LongBeach Dataset		
IG	11	thal, fbs, trestbps, cp, sex, chol, restecg, ca, thalach, slope, oldpeak
GS	1	exang
Switzerland Dataset		
IG	4	thal, fbs ,trestbps, cp
GS	1	exang

Validation is based on the number of correctly classified instances. It is observed from this figure that GS feature selection technique has not improved the accuracy for any of the classifiers for the Hungarian dataset, but there is significant improvement when IG feature selection technique is applied. This implies that all the 13 attributes are prominent and feature reduction is not significant. For the LongBeach dataset, IG feature selection technique has a performance dip only for LR, which means that these attributes are leading to underfitting in dataset, which can be further validated in stage 2. A similar performance dip is observed only in SVM in GS for the Switzerland dataset, this is due to the sensitiveness of the classifier to the kernel function used. However, keeping these small variations of results in mind, it is found that even though the Cleveland and Hungarian datasets are uniformly distributed an accuracy ranging from 60-70% is observed, compared to an accuracy of 30-35% for the LongBeach and Switzerland datasets which have skewed distribution. Thus, it reveals that even though feature selection techniques were applied, there is no significant improvement in overall performance of the classifiers. Thus 13 features with 5 classes is not desirable for making effective decisions on the risk level of cardiovascular disease, there is a need to study the remaining features and their prominence in effective decision making.

The second stage of analysis was done using the 75 attributes which includes the blood vessels of the heart i.e. left main trunk (*lmt*), left anterior descending artery (*ladprox*, *laddist*), right coronary artery (*rcaprox*, *rcadist*), cxmain, om1, lvx4 etc., and five classes C0-C4. The UCI repository can be referred for a detailed description of these attributes. Table 5 shows the number of attributes and their names for each of the datasets after the feature selection techniques were applied.

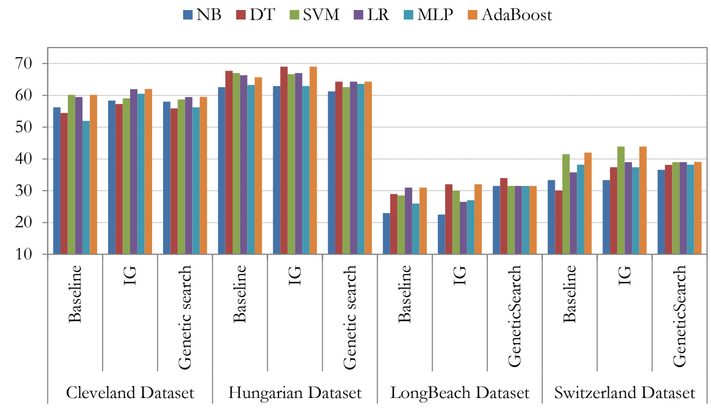


Figure 2. Classifier performance of stage 1 analysis

It is observed from this table that for the LongBeach and Switzerland datasets the number of attributes selected after both the feature selection techniques are same. Similar to stage 1 validation is done for the classifiers and results were reported in Figure 3.

Table 5. Selected attributes from 75 attributes using IG and GS

Feature Selection Technique	# of selected attributeses	List of attributes
Clevland Dataset		
IG	8	laddist, om1, rcaprox, cxmain, lmt, thal, rcadist, ladprox
GS	11	cp, oldpeak, ca, thal, lmt, ladprox, laddist, cxmain,om1, rcaprox, rcadist
Hungarian Dataset		
IG	15	rcaprox, cxmain, ladprox, lmt, slope, exang, cp, oldpeak, painexer (pain during exertion), laddist, thaltime, relrest (relieved after test), lvx4, rcadist, om1
GS	7	painexe, slope, lmt, ladprox, laddist, cxmain, rcaprox
LongBeach Dataset		
IG/GS	8	rcaprox, lmt, cxmain, ladprox, om1, lvx4, laddist, rcadist
Switzerland Dataset		
IG/GS	7	lmt, rcaprox, ramux, ladprox, cxmain, om1, ladddist

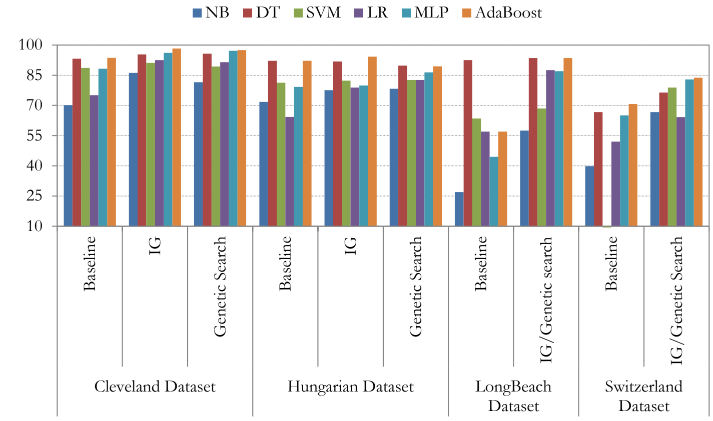


Figure 3. Classifier performance of stage 2 analysis

Result Analysis

Significant observation reveals that Adaboost with DT as classifier has the best performance across all the four datasets with a maximum accuracy of 98%. Next is DT, which shows good performance even with skewed data distribution of the LongBeach dataset and the Switzerland dataset. Similarly all other classifiers show significant improvement in performance. Highlights of analysis are: NB has maximum performance of about 86% compared to the other classifiers which have a performance of at least 90% for the Cleveland and Hungarian datasets, making it an average performing classifier. This indicates high dependence between the attributes. In the LongBeach and Switzerland datasets, NB has a maximum performance of 66%. SVM is giving a performance of 80-90% for the Cleveland and Hungarian datasets while an accuracy of 68% and 78% were obtained for LongBeach and Switzerland datasets. SVM is natively known to work well with binary class, since the experiment was conducted using five classes (C0-C4) there is variation in its performance across different datasets. LR gives an accuracy of 64% in Switzerland dataset which indicates underfittingdata. From Table 5 it can be seen that for the LongBeach dataset using LR, there is no attribute in common with the 13 attributes used in stage1, so that means that the attributes leading to underfitting in stage1 are not present in stage2 and thus the performance has improved significantly. Similarly a performance variation is observed in MLP for the Switzerland dataset which is due to its skewed distribution. There are few attributes like *chest pain(cp), slope, oldpeak , thal* etc. which are part of the 13 attributes and are found in stage 2 analysis used for the Cleveland dataset and the Hungarian dataset, while the LongBeach and Switzerland dataset do not have any attributes from stage1. Keeping this mind when Figure3 is studied, it is observed the Cleveland dataset and the Hungarian dataset have a maximum accuracy of 98% and 97% respectively. Though there is no attributes in common from stage2, the Longbeach dataset and the Switzerland dataset have a maximum performance of 93% and 84% respectively. This gives an understanding about the actual relevance of these 13 attributes.

This has led to the need to study the common features across all datasets and its prominence in understanding the risk level of cardiovascular disease which is shown in Figure 4.

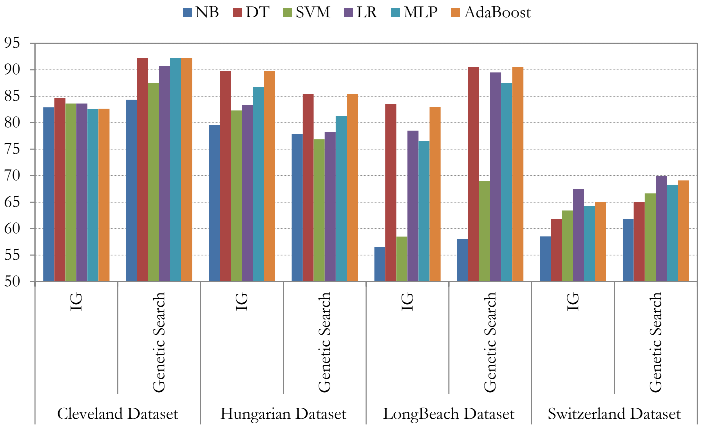


Figure 4. Performance of the common attributes from stage 2 across all datasets

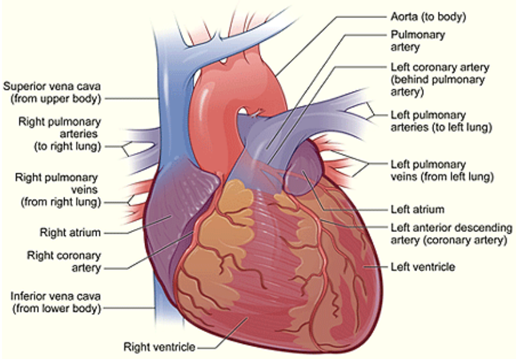
It was observed within the classes that there has been a deal of confusion during classification. A snapshot validating this is shown in Figure 5a) and Figure 5b), where out of 31 instances for C2 in Figure 5b), 26 were correctly classified, 3 were incorrectly classified as c1 and 2 incorrectly classified as C3. Similar observation in Figure 5a) for classes C1, C2, C3. But when all the cases were analyzed, it was seen that this confusion was mainly observed in classes 2 and 3. This may be due to the improper data distribution for this class within the datasets. This is one of the reasons why only a maximum performance of 98% is observed. Furthermore, to get a clear understanding of the prominence of the reduced set of features another study was done, where common attributes for each of the feature selection technique were analyzed and these features were studied for all the classifiers across all datasets and represented in Figure 4.

	c0	c1	c2	c3	c4	**classified as**
	187	1	0	0	0	**c0=0**
	0	32	3	2	0	**c1=1**
	0	5	17	4	0	**c2=2**
	0	0	8	20	0	**c3=3**
a)	0	0	0	0	15	**c4=4**

	c0	c1	c2	c3	c4	**classified as**
	157	0	0	0	0	**c0=0**
	0	50	0	0	0	**c1=1**
	0	3	26	2	0	**c2=2**
	0	0	0	32	0	**c3=3**
b)	0	0	0	0	11	**c4=4**

Figure 5. a) Confusion matrix of Cleveland dataset using IG and Adaboost classifier yielding an accuracy of 98.22%; **b)**Confusion matrix of Hungarian dataset using IG and Adaboost classifier yielding an accuracy of 92.17%

The common attributes found are *laddist, ladprox, rcaprox, lmt, cxmain* for GS, it is the same for IG with an additional attribute *om1*. As observed in Figure 5 the common attributes which are the major blood vessels of the heart, are giving an accuracy of nearly 92% for DT and an average performance of 60% is observed across all classifiers for all datasets. Thus, these common features have greater contribution for diagnosis of cardiovascular disease. The next step is to understand the anatomic relevance of these attributes. These attributes are the major blood vessels of the heart as shown in Figure 6. They play a major role in understanding the risk level of cardiovascular diseases and can be analyzed and studied using the coronary angiography undertaken by the medical practitioner.

Attribute Name	**Anatomic Name**	
Laddist	Left Anterior descending artery distance	
Ladprox	Left Anterior descending artery Proximation	
Rcaprox	Right Coronary artery proximation	
Lmt	Left Main Trunk	
Cxmain	Blood vessel	
Om1	Blood vessel	

Aorta (to body)
Pulmonary artery
Left coronary artery (behind pulmonary artery)
Superior vena cava (from upper body)
Right pulmonary arteries (to right lung)
Left pulmonary arteries (to left lung)
Right pulmonary veins (from right lung)
Left pulmonary veins (from left lung)
Left atrium
Right atrium
Left anterior descending artery (coronary artery)
Right coronary artery
Left ventricle
Inferior vena cava (from lower body)
Right ventricle

Figure 6. Anatomic relevance of attributes [23]

Conclusion and Future Work

The analysis was carried out in two stages, using different numbers of attributes and five classes. It is understood that 13 attributes are not sufficient to understand the risk level of the cardiovascular disease. Analysis of 75 attributes using the feature selection techniques has yielded significant and relevant information and improved the accuracy of the classifiers. It was found that among the classifiers Adaboost with DT as classifier has the maximum performance i.e. 98% for Cleveland dataset and for Switzerland and LongBeach dataset which have skewed distribution a maximum performance of almost 90% is observed. Another significant observation is about the class 2 and class 3 which appears to be confused due to improper data distribution and affects the accuracy of the classifiers. From the various stages of analysis undertaken, it is understood that Machine Learning techniques have derived attributes which are of anatomic relevance. Thus they can yield significant and relevant conclusions for clinical data.

In future, association rules can be derived which help in understanding the relationship between the attributes and derive meaningful conclusions, which will give a better understanding about the risk level of the cardiovascular disease. This analysis can be done on various other domains of healthcare and effective decision making systems can be introduced to provide quality and cost effective healthcare for society.

References

1. Hand D, Mannila H, Smyth P, Principles of data mining. Bradford book. Cambridge, Massachusetts, London: The MIT Press; 2001.
2. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 2012;13:395-405.
3. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, et al. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. Springer Science+Business Media, LLC 2011.
4. NHS Choices. Coronary Heart Disease [Internet]. 2014 [updated 2014 Sept 26; cited 2015 Jan 10]. Available from: http://www.nhs.uk/Conditions/Coronary-heart-disease/pages/Introduction.aspx
5. Duraisamy K, Haridass K. An Effective Comparison of SVM and CN2Rule Using Heart Dataset: A Survey. International Journal of Advanced Research in Computer and Communication Engineering 2014;3:5774-5776
6. Sudhakar K, Manimekalai. M. Study of Heart Disease Prediction using Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering. January 2014;4:1157-1160.
7. Aditya Sundar N, Pushpa Latha P, Rama Chandra M. Performance Analysis of Classification Data Mining Techniques Over Heart Disease Database. IJESAT 2012;2(3):470-478.
8. Detran R. Heart Disease Dataset [Internet]. 1988 [updated 1988 July; cited 2015 Jan 10]. Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/cleveland.data
9. Janosi A. Heart Disease Dataset [Internet]. 1988 [updated 1988 July; cited 2015 Jan 10]. Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/hungarian. data
10. Aha DW. Heart Disease Dataset[Internet]. 1988 [updated 1988 July; cited 2015 Jan 10]. Available from: http://archive.ics.uci. edu/ml/machine-learning-databases/heart-disease/long-beachv.data
11. Steinbrunn W. Heart Disease Dataset [Internet]. 1988 [updated 1988 July; cited 2015 Jan 10]. Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/switzerland.data
12. Can M. Diagnosis of Cardiovascular Diseases by Boosted Neural Networks. Southeast Journal of Soft Computing 2013;2:91-102.
13. Sunila Prabhat, Panday Nirmal Godara. Decision Support System for Cardiovascular Heart Disease Diagnosis using Improved Multilayer Perceptron. International Journal of Computer Applications 2012;45:12-20.
14. Noor Akhmad Setiawan, Venkatachalam PA, Ahmad Fadzil M Hani. Diagnosis of Coronary Artery Disease using Artificial Intelligence Based Decision Support System. Proceedings of the International Conference on Man-Machine Systems (ICoMMS), October 11-13, 2009.
15. Shruti Ratnakar, Rajeshwari K, Jacob R. Prediction of Heart Disease using Genetic Algorithm For Selection of Optimal Reduced Set Of Attributes, International Journal of Advanced Computational Engineering and Networking 2013;1:51-55.
16. Nidhi Bhatla Kiran Jyoti. A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic. International Journal of Computer Applications 2012;54(17):16-21.
17. Tang Chun Wong, Genetic Algorithms[Internet] 1996 [updated 1996 June 16; cited 2015 Jan 10]. Available from: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/tcw2/report.html
18. Seo Y-W. Class InfoGain [Internet]. 2003 [updated 2003 August 15; cited 2015 Jan 10]. Available from: http://www.cs.cmu.edu/~youngwoo/projects/textminer/textminer-docs/doc/textminer/featureselection/InfoGain.html
19. Mitchell T. Machine Learning. McGraw Hill; 1997.
20. Harrington P. Machine Learning in Action. New York: Manning Publication, Special Sales Department; 2012.
21. Machine Learning Group at University of Waikato. Downloading and Installing Weka [Internet] [cited 2015 Jan 10] Available from: http://www.cs.waikato.ac.nz/ml/weka/downloading.html
22. WebMD. Heart Disease Health Center [Internet] 1999 [cited 2015 January 10]. Available from: www.webmd.com/heart-disease/guide/heart-disease-diagnosis-tests
23. Nhlbi, Nih. Anatomy of the Heart. 2011 [updated 2011 November 17; cited 2015 January 10]. Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/hhw/anatomy