Expert Systems with Applications 40 (2013) 96–104

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Computational intelligence for heart disease diagnosis: A medical knowledge driven approach

Jesmin Nahar [a,*], Tasadduq Imam [a], Kevin S. Tickle [a], Yi-Ping Phoebe Chen [b]

[a] Faculty of Arts, Business, Informatics and Education, Central Queensland University, Queensland, Australia
[b] Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia

ARTICLE INFO

Keywords:
Cleveland data
Heart disease
Computational intelligence
Classification
Feature selection

ABSTRACT

This paper investigates a number of computational intelligence techniques in the detection of heart disease. Particularly, comparison of six well known classifiers for the well used Cleveland data is performed. Further, this paper highlights the potential of an expert judgment based (i.e., medical knowledge driven) feature selection process (termed as MFS), and compare against the generally employed computational intelligence based feature selection mechanism. Also, this article recognizes that the publicly available Cleveland data becomes imbalanced when considering binary classification. Performance of classifiers, and also the potential of MFS are investigated considering this imbalanced data issue. The experimental results demonstrate that the use of MFS noticeably improved the performance, especially in terms of accuracy, for most of the classifiers considered and for majority of the datasets (generated by converting the Cleveland dataset for binary classification). MFS combined with the computerized feature selection process (CFS) has also been investigated and showed encouraging results particularly for NaiveBayes, IBK and SMO. In summary, the medical knowledge based feature selection method has shown promise for use in heart disease diagnostics.

1. Introduction

Various classification and regression processes have been used to identify heart disease (Boors et al., 2000; Das, Turkoglu, & Sengur, 2009; Detrano et al., 1989; El-hanjouri, Alkhaldi, Hamdy, & Alim, 2002; Skalak, 1997). In particular, focus has been made on the University of California Irvine (UCI) heart disease dataset (also known as the Cleveland dataset (Uci, 2009) and different computational intelligence algorithms have been used. But, existing investigations are, to the best of the author's knowledge, yet to show a comparative research that considers modern classification techniques and imbalanced nature of the data, and employs a feature selection process which incorporates medical knowledge. Medical knowledge is important for feature selection in this area since computer automated process may remove features important or select feature that are less likely related from clinical view. The research presented in this paper highlights this issue and the findings may aid in contributing to future identification of heart disease.

The plan of this paper is as follows: Section 2 provides an overview of existing research on using computational intelligence techniques in heart diseases diagnosis. Section 3 details the datasets and 4 demonstrate the experimental setup that has been used in this research. Section 5 then presents the comparative research of the different classification algorithms and describes the best suited ones for this problem. Section 6 describes significant risk factors for heart disease from medical point of view. Section 7 presents the results of the comparison of computer feature selection (CFS) process and medical knowledge based feature selection for heart disease dataset. Section 8 proposes the medical knowledge motivated feature selection (MFS), as well as a process combining CFS with MFS. Finally, Section 9 concludes the paper with a summary of findings and future research directions.

2. Computational intelligence for heart disease diagnostics

This section provides an overview of existing research on using computational intelligence techniques in heart diseases diagnosis and points to the limitations that motivated this research. Cardiovascular disease is a highly mortal disease with over 17 million deaths globally (Smith, 2010). So, early detection and treatment of the disease are imperative. Researchers have used different computational intelligence techniques to improve heart disease diagnostics over the years. A particular heart disease diagnostic dataset widely popular with data mining researchers is the publicly available University of California Irvine, Cleveland dataset (Uci, 2009). Some of the key researches on this datasets are:

* Corresponding author. Tel.: +61 07 40232112; fax: +61 07 49309700.
E-mail addresses: j.nahar@cqu.edu.au (J. Nahar), t.imam@cqu.edu.au (T. Imam), k.tickle@cqu.edu.au (K.S. Tickle), phoebe.chen@latrobe.edu.au (Y.-P.P. Chen).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.07.032

- Aha & Kibler (1988) used the dataset to predict effectiveness of instance-based algorithms and achieved 77% and 74.8% accuracy for NTgrowth and C4.5 techniques.
- Detrano et al. (1989) investigated a probabilistic algorithm to diagnose the risk of coronary artery disease and concluded that patients experiencing chest pain and transitional disease occurrences are the higher risk subjects.
- Gennari, Langley, & Fisher (1989) explored a conceptual clustering system and gained an acceptable accuracy (78.9%).
- Edmonds (2005) worked on the Cleveland data set with focus on comparing global evolutionary computation approaches, and observed some prediction performance improvement with a new approach. However, performance of the proposed technique is dependent on the attributes selected by the algorithm.

Other than these works, several researches have focused on diverse aspects of heart disease diagnosis on different datasets (Avci, 2009; Boors et al., 2000; Doyle, Temko, Marnane, Lightbody, & Boylan, 2010; El-hanjouri et al., 2002; Gamboa, Mendoza, Orozco, VARGAS, & Gress, 2006; Maglogiannis, Loukis, Zafiropoulos, & Stasis, 2009; Obayya & Abou-chadi, 2008; Zheng, Jiang, & Yan, 2006; Kim, Lee, Cho, & Oh, 2008). Also, different researchers have used different machine learning techniques in related research. These include: fuzzy support vector clustering for the identification of heart disease (Gamboa et al., 2006), prototype development using data mining techniques, mainly decision trees, Naive Bayes and Neural Networks, (Palaniappan & Awang, 2008) diagnostic system improved using feature extraction and Hidden Markov Models (HMM) (El-hanjouri et al., 2002), a data fusion approach recommended for classifying heart diseases (Obayya & Abou-chadi, 2008), an intelligent system based on genetic-support vector machines (GSVM) (Avci, 2009), use of an automated detection system based on the SVM classification (Maglogiannis et al., 2009), a committee machine (CM) based on an ensemble of Multilayer Perceptions (MLP) (Zheng et al., 2006), a computerized cardiovascular disease diagnosis and categorization system (Kim et al., 2008) and decision trees and SVM to predict heart disease (Soman, Shyam, & Madhavdas, 2003).

Feature selection has also been applied in heart disease diagnostics, but for mainly datasets other than Cleveland. For instance, Zhao, Chen, Hou, Zheng, & Wang (2010) used backward elimination procedure along with a novel algorithm, Fan & Chaovalitwongse (2010) suggested a novel optimization framework for getting improved feature selection in classification. Several other researchers have also noted impact of feature selection in different heart disease diagnosis (Chang, 2010; Hanbay, 2009; Qazi et al., 2007; Zhao, Guo et al., 2010). Further, feature selection processes have often been found to improve the prediction performance of different classifiers (Abraham, Simha, & Iyengar, 2007; Cheng, Wei, & Tseng, 2006; Devaney & Ram, 1997; Polat & Guenes, 2009; Sethi & Jain, 2010; Wang & Ma, 2009; Zhao, Chen et al., 2010).

It is observed that a number of different classifier have been used to diagnose heart disease in the different studies. The comparison of different algorithms in order to identify the heart disease, however, has to date not received appropriate focus. In addition, the literature has not taken into account medical knowledge based feature selection for medical datasets during the classification of heart disease. Computer based feature selection (CFS) selects features randomly, through calculating the significance of the attributes and by considering the individual predictive capacity. So, there is a chance to discard medically important factors for a specific disease. For instance, as shown in Fig. 4, applying computerized feature selection (CFS) on Cleveland dataset (with healthy as the positive class) discards medically established attributes like age, cholesterol, fasting blood sugar, resting blood pressure and ECG characteristics. This sort of outcomes is doubted by medical practitioners and reduces the significance of the automated system. So, a feature selection process motivated by medical knowledge is important.

The literature also indicates that in most cases complex and time intensive algorithms have been recommended. Well-known standard classification algorithms are, however, more easily accessible due to its availability in different software packages. From the medical practitioner's point of view, in particular, a comprehensive analysis of well-established classifiers is, so, of essence.

This research focuses on these issues. As Cleveland dataset is considered a benchmark data in many existing research, this research also uses this dataset. The study provides a comparative suitability of commonly used classifiers. In addition, the research investigates medical knowledge guided feature selection process for classification of heart disease.

3. Dataset details

As mentioned earlier, the popular and publicly available UCI heart disease dataset is used in this research. The UCI heart disease dataset consists of a total 76 attributes. However, majority of the existing studies have used only a maximum of 14 attributes (Uci, 2009; Uci, 2010). Different datasets have been based on the UCI heart disease data. Computational intelligence researchers, however, have mainly used the Cleveland dataset consisting of 14 attributes. The 14 attributes of the Cleveland dataset along with the values and data types are as follow (Uci, 2009; Uci, 2010).

1. Age: age in years (*numeric*);
2. Sex: male, female (*nominal*);
3. Chest pain type (CP): (a) typical angina (angina), (b) atypical angina (abnang), (c) non-anginal pain (notang), (d) asymptomatic (asympt) (*nominal*). From medical point of view,
   (a) Typical 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 (Baliga & Eagle, 2008; Diagnosis, 2010; Kaul, 2010).
   (b) Atypical angina refers to the condition that the patient's symptoms are not detailed and so the probability of blockages is lower (Baliga & Eagle, 2008; Diagnosis, 2010; Kaul, 2010).
   (c) Non-angina pain is the stabbing or knife-like, prolonged, dull, or painful condition that can last for short or long periods of time (Diagnosis, 2010; Mengel & Schwiebert, 2005; Society, 1945).
   (d) Asymptomatic pain shows no symptoms of illness or disease and possibly will not cause or exhibit disease symptoms (Pickett, 2000; Freedc, 2010);
4. Trestbps: patient's resting blood pressure in mm Hg at the time of admission to the hospital (*numeric*);
5. Chol: Serum cholesterol in mg/dl;
6. Fbs: Boolean measure indicating whether fasting blood sugar is greater than 120 mg/dl: (1 = True; 0 = false) (*nominal*);
7. Restecg: electrocardiographic results during rest. Three types of values normal (norm), abnormal (abn): having ST-T wave abnormality, ventricular hypertrophy (hyp) (*nominal*);
8. Thalach: maximum heart rate attained (*numeric*);
9. Exang: Boolean measure indicating whether exercise induced angina has occurred: 1 = yes, 0 = no (*nominal*);
10. Oldpeak: ST depression brought about by exercise relative to rest (*numeric*);
11. Slope: the slope of the ST segment for peak exercise. Three types of values upsloping, flat, downsloping (*nominal*);

Table 1
The heart-disease datasets (Uci, 2009).

Dataset name	Class label considered as positive	No. of positive class instances	No. of negative class instances	Class (Label: No. of instances)		Status indicated by positive class
H-O	0 (healthy)	165	138	1:165	2:138	Healthy, Sick
Sick-1	1 (sick1)	56	247	1:56	2:247	Sick1
Sick-2	2 (sick2)	37	266	1:37	2:266	Sick2
Sick-3	3 (sick3)	36	267	1:36	2:267	Sick2
Sick-4	4 (sick4)	14	289	1:14	2:289	Sick4

12. Ca: number of major vessels (0–3) colored by fluoroscopy (*numeric*);
13. Thal: the heart status (normal, fixed defect, reversible defect) (*nominal*);
14. The class attributes: value is either healthy or heart disease (sick type: 1, 2, 3, and 4).

The dataset has five class attributes indicating either healthy or one of four sick types. For this research, multi-class classification problem is converted into a binary classification problem. The reason for this is SMO, a robust and modern algorithm and used in the experiments, is principally a binary classifier. For the conversion into binary, one of the class labels was considered as positive and the rest as negative. This way five datasets were created. Table 1 shows the characteristics of these datasets. The generated datasets are referenced using the symbols: H-0 (healthy), Sick1, Sick2, Sick3 and Sick4 respectively.

4. Research design

To provide a comparison among the well popular classification algorithms, four performance metrics were used in our experiment. These are: accuracy, true positive rate (TP), F-measure, and time. Here, accuracy was the overall prediction accuracy, true positive rate (TP) was the accurate classification rate for the positive classes, and F-measure indicates the effectiveness of an algorithm when the accurate prediction rates for both of the classes are considered. Also, training time was considered to compare the computational complexity for learning.

In the case of medical data diagnosis, many researchers have used a 10-fold cross validation on the total data and reported the result for disease detection, while other researchers have not used this method for heart disease prediction (Abdel-aal, 2005; Baek, Tsai, & Chen, 2009; Chen et al., 2007; Dash, 2008; Fountoulaki, Karacapilidis, & Manatakis, 2010; Kumar & Shelokar, 2008; Mei, Ma, Ashley-koch, & Martin, 2005; Polat & Güne, 2007; Xing, Wang, Zhao, & Gao, 2007). We argue that selecting the best training parameters on a validation set and reporting prediction on a test set is more authentic than simply performing a 10-fold cross validation on a training set. However, to relate with common culture, we have used both the train-test split method and 10-fold cross validation when comparing the algorithms.

The experimental software used was Kumar & Shelokar, 2008; Witten & Frank, 2005). Five classification techniques, as implemented in Weka, were used. These are: (accuracy, TP, F-measure, and time).

5. Comparison of algorithms

Weka provides facilities to report performance of classifiers by performing a 10-fold cross validation on a provided dataset and report performance results on the given dataset. This is the method generally used by many researches. However, this method is expected to be biased to the training data and may not reflect the expected performance when applied on real-life data. So, in addition to generally used 10-fold cross validation, we have also performed a train-test split on the dataset and then used a 10-fold cross validation to select the best parameter for training. Performance results were presented based on the prediction outcomes of the test set. For each of the datasets, a stratified sampling process was used to select two-thirds of the data for training and the rest for prediction. The CVParameter selection tool provided by Weka was used for the train-test split.

The results obtained using the two experimental processes are shown in Table 2. The following discussion refers to these two experiments as the 10-fold and the CVP 10-fold.

From the results (Table 2), it was found that for the 10-fold, SMO is the best performing algorithm for all the datasets in terms of accuracy. But the CVP 10-fold results showed that the SMO algorithm is the best for three datasets only (sick-1, sick-2, and sick-4). For the H-O (healthy) dataset, SMO shows better performance in terms of TP (0.891) and F-measure (0.862) compared to other algorithms using the10-fold cross validation. But using the CVP 10-fold, the results showed that Naive Bayes was best in terms of accuracy, IBK was best in terms of TP and AdaBoostM1 was best in terms of F-measure.

It is observed that the two experimental processes gave varied results. But, in terms of accuracy, the best performing algorithm was the SMO for both the 10-fold and CVP 10-fold. Thus it can be concluded that when considering accuracy as the key performance measure, the experimental result showed that SMO is the best suited classification algorithm among six algorithms for the UCI heart disease dataset. In respect to performance metrics like TP and F-measure, no such clear outcome exists and the choice of best algorithm would then be dependent upon the characteristics of the dataset.

6. Responsible risk factors for heart disease from a medical point of view

Single risk factors are not notably sensitive to make out all individuals at high risk of heart disease. There are different factors of heart disease, although none of each of these factors are less important for disease diagnosis however, some of those factors deserve to be more actively considered when diagnosing heart disease according to some medical literature; emphasizing that certain factors need to be considered during heart disease diagnosis. For example in heart disease diagnosis, most of the time exercise stress testing is an important factor in assessing the coronary heart disease. The literature argues that this test is not necessarily a wise decision for the majority of patients with a recognized unstable angina (type of chest pain) condition or those with an abnormal ECG. One could argue that heart disease diagnosis, if the patient is able to exercise for 6 or 8 min without any complaints in the chest that patient could be considered to possess a healthy heart condition. In contrast, if patients display any pain or discomfort in the chest during exercise, or if their heart rate falls or goes up, this could suggest an abnormal condition of the heart;

Table 2
Performance table for 10-fold and CVP 10-fold. (Bold values indicate the best algorithm. Performance close to the best algorithm is also bold. The column 10-fold shows the results from the 10-fold cross validation, while CVP 10-Fold shows the results from the train-test split and 10-fold cross-validations being applied on a training set in order to choose the best learning parameters.)

Dataset	Algorithms	Accuracy (%)		TP		F-measure		Training time (s)	
		10-fold	CVP-10-fold	10-fold	CVP-10-fold	10-fold	CVP-10-fold	10-fold	CVP-10-fold
H-O	Naive Bayes	83.83	**81.88**	0.797	0.782	0.818	0.819	0.02	0.02
	SMO	**84.49**	75.25	**0.891**	0.764	**0.862**	0.771	0.14	0.14
	IBK	76.90	76.24	0.80 0	**0.855**	0.79	0.797	**0**	**0**
	AdaBoostM1	83.50	81.19	0.855	0.818	0.849	**0.826**	0.03	0.03
	J48	76.57	76.24	0.806	0.745	0.789	0.774	0.03	0.03
	PART	81.52	79.21	0.830	0.818	0.830	0.811	0.03	0.02
Sick-1	Naive Bayes	74.92	72.28	0.196	0.111	0.224	0.125	0.02	**0**
	SMO	**81.52**	**82.18**	0	0	0	0	0.14	0.16
	IBK	73.27	72.28	**0.321**	0.111	**0.308**	0.125	**0**	0.02
	AdaBoostM1	81.52	**82.18**	0	0	0	0	0.03	0.05
	J48	81.52	**82.18**	0	0	0	0	0.03	0.03
	PART	75.25	71.29	0.214	**0.222**	0.242	**0.216**	0.03	0.02
Sick-2	Naive Bayes	78.55	79.21	**0.405**	**0.333**	**0.772**	**0.276**	0.02	0.02
	SMO	**87.79**	**88.12**	0	0	0	0	0.16	0.16
	IBK	82.84	79.21	0.216	0.167	0.586	0.16	**0**	**0**
	AdaBoostM1	86.80	86.14	0	0	0	0	0.03	0.03
	J48	86.14	86.14	0	0.167	0	0.222	0.03	0.02
	PART	79.87	85.15	0.162	0.167	0.539	0.211	0.02	0.02
Sick-3	Naive Bayes	81.52	81.19	**0.472**	**0.583**	**0.378**	**0.424**	**0**	0
	SMO	**87.88**	88.11	0	0	0	0	0.22	0.16
	IBK	83.17	87.13	0.194	0.333	0.215	0.381	0	**0**
	AdaBoostM1	87.46	86.14	0.056	0.083	0.095	0.125	0.03	0.03
	J48	87.46	**88.12**	0.083	0	0.136	0	0.05	0.02
	PART	82.18	83.17	0.167	0	0.182	0	**0**	0.03
Sick-4	Naive Bayes	89.77	92.08	0.143	0	0.114	0	**0**	0
	SMO	**95.38**	**96.04**	0	0	0	0	0.19	0.16
	IBK	92.74	**95.05**	**0.214**	0	**0.644**	0	**0**	**0**
	AdaBoostM1	**95.38**	**96.04**	0	0	0	0	0.03	0.03
	J48	**95.38**	**96.04**	0	0	0	0	**0**	0.02
	PART	93.40	**96.04**	0.143	0	0.167	0	**0**	0.02

or when the patient is resting or exercising his/or her ECG showing an abnormal reading could be an indication of a heart disease status. A positive exercise stress test is specified by abnormal horizontal or downsloping ST segment depression (Khan, 2005). Those factors might support the diagnosis of heart disease more strongly than other factors and are discussed below.

From this overview of medical literature (details in Section 6 and Table 3), it can be seen that factors such as cholesterol, hypertension (blood pressure), heart rate, resting ECG, blood sugar, diabetes, exercise induced angina, stress and older age, are the most significant factors for predicting heart disease. Several books and articles mentioned those particular factors being effective and therefore should be considered during disease diagnosis. In the UCI Cleveland heart disease dataset, there are eight factors of medical significance that are considered for feature selection as MFS. The factors are: age, chest pain type (angina, abnang, notang, asympt), resting blood pressure, cholesterol, fasting blood sugar, resting heart rate (normal, abnormal, ventricular hypertrophy), maximum heart rate, and exercise induced angina. As discussed previously, feature selection based on medical knowledge is an important factor in heart disease diagnosis. It can be argued that during feature selection, if significant symptoms related to heart disease are not considered then there is a strong likelihood that the diagnosis runs the risk of neglecting the most important factors. In the subsequent experiment, the knowledge derived from the medical literature survey is taken into account and the eight factors mentioned above are used for feature selection.

7. Comparison of automated and medical knowledge based feature selection

This section presents a comparison between medical knowledge based feature selection and computerized feature selection. For the computerized feature selection process, CfsSubsetEval attribute selection (using BestFirst search strategy) provided by Weka (Witten & Frank, 2005) was used. In later discussions, the symbol CFS (computer feature selection) is used to indicate this process. In Fig. 1, the attributes selected by MFS and CFS for H-O dataset are shown below.

CFS computational modeling selection is based on the significant predictor and as a consequence does not necessarily consider medically significant factors. It can be seen that medically important attributes such as age, resting blood pressure, cholesterol, fasting blood sugar and resting ECG have been discarded by CFS for healthy (H-O) dataset. In sick-1 (Fig. 2) age, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, max heart rate were not considered relevant by CFS. Similarly for sick-2 (Fig. 3), age, resting blood pressure, cholesterol, resting ECG was discarded by CFS. While for sick-3 (Fig. 4) age, resting blood pressure, cholesterol, resting ECG, max heart rate has been discarded by CFS.

And for sick-4 (Fig. 5), age, resting blood pressure, cholesterol, fasting blood sugar, max heart rate and exercise induced angina were not considered. For sick-4 only four attributes were selected by CFS: chest pain type, resting ECG, slope, number of vessels colored and thal (heart status). Factors such as age, resting blood pressure, cholesterol, maximum heart rate, exercise induced angina, fasting blood sugar, selected by MFS, were disregarded by CFS. The results indicate that the CFS method selected attributes with the significance association among the factors. The six classification algorithms were executed on each of the datasets and feature selection was applied based on CFS or MFS. Table 4 presents performance of the algorithms using datasets selected on the basis of MFS, CFS and CVP-10 fold results. Performances are quantified in terms of accuracy, TP and F-measure.

Results indicated that for H-O, MFS performance improved when compared with CFS in terms of accuracy for two cases

Table 3
Medically important factors related to heart disease.

Source	Considered Factors														
	Age	Angina	Blood pressure	Blood sugar	Chest pain	Cholestoerol	Diabetes	ECG	Exercise induced angina	Fasting serum glucose & total serum cholesterol	Heart rate	Hypertension	Low exercise workload	Smoking	Stress
Boyko et al., 2004				✓											
Dietrich et al. 2008	✓		✓	✓		✓									
Ding et al., 2004			✓												
Edlin and Golanty, 2009	✓	✓		✓	✓	✓					✓	✓			
Edlin et al., 1999	✓		✓	✓		✓									
Facchini et al. 2009			✓			✓									
Fuster et al., 2005	✓		✓	✓	✓	✓		✓	✓		✓				
Goodpaster et al., 2003				✓											
Hales, 2008	✓		✓	✓		✓									
Hales, 2009	✓		✓	✓		✓									
Hayashi et al., 2004			✓												
Hoegerand Hoeger, 2010	✓			✓	✓	✓		✓			✓	✓			
Huikuri2009								✓			✓				
Kanaya et al., 2004				✓											
Khang et al. 2008			✓							✓					
Kurl et al. 2009													✓		
Lindeberg et al. 2009	✓		✓			✓									
Lindsay and Gaw, 2004	✓		✓	✓	✓	✓		✓	✓		✓				
Lyerly et al. 2008								✓							
Maylunas and Mironenko, 2005	✓		✓				✓	✓	✓		✓				
Miller, 2008	✓			✓	✓	✓					✓	✓			
Mittal, 2005	✓		✓	✓	✓	✓		✓	✓		✓				
Mozaffiman et al. 2008	✓			✓								✓			
Nyman et al. 2009								✓							
Peacock et al., 2009			✓	✓	✓	✓		✓	✓		✓				
Schenck-Gustafison 2009	✓			✓								✓			✓
Sizer and Whitney, 2007	✓		✓	✓		✓									
Stansfeld et al. 2009					✓										
Tan et al. 2008	✓														
Yusuf et al. 2004				✓								✓		✓	✓

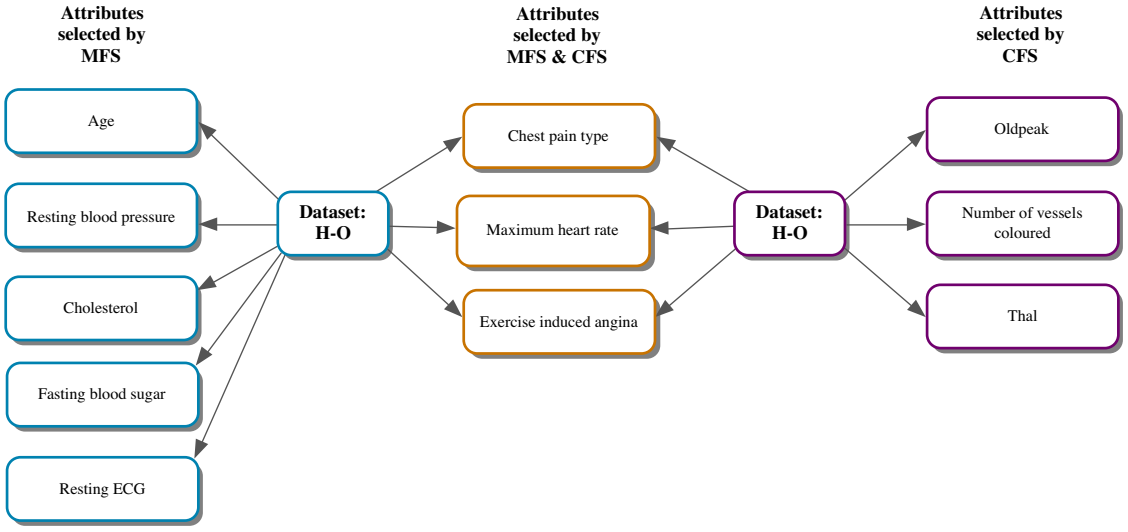

Fig. 1. Attributes selected by MFS and CFS for dataset (H-O).

(IBK-100, PART-86.77) and comparable to CFS (J48-80.88), the performance of MFS was comparable to CFS for two cases (SMO-82.35, J48-82.35) and better than CFS (Naive Bayes-85.29, AdaBoostM1-82.35) in 2 cases. Similar characteristics were observed for sick-2,

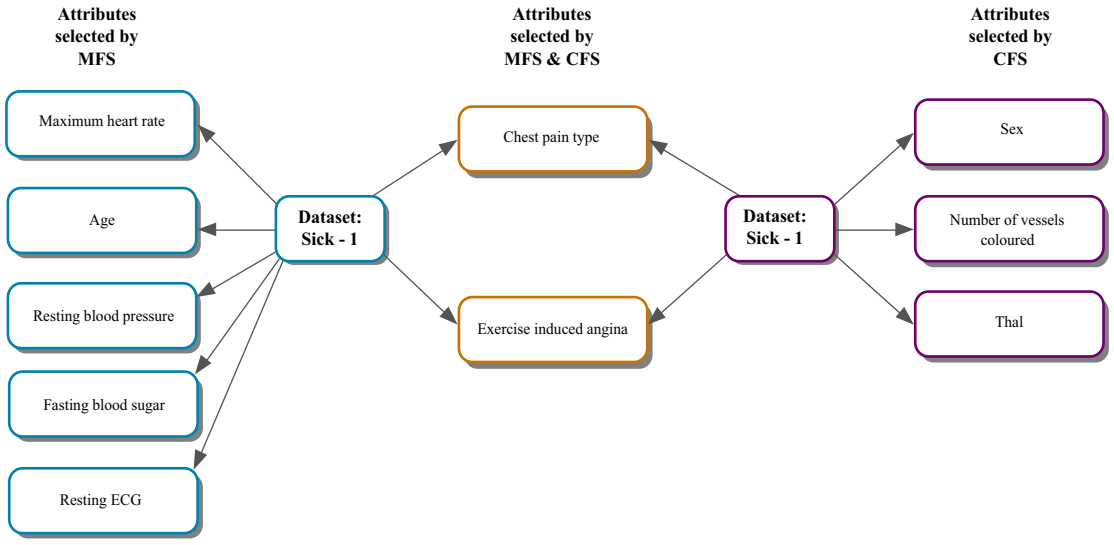

Fig. 2. Attributes selected by MFS and CFS for dataset (sick-1).

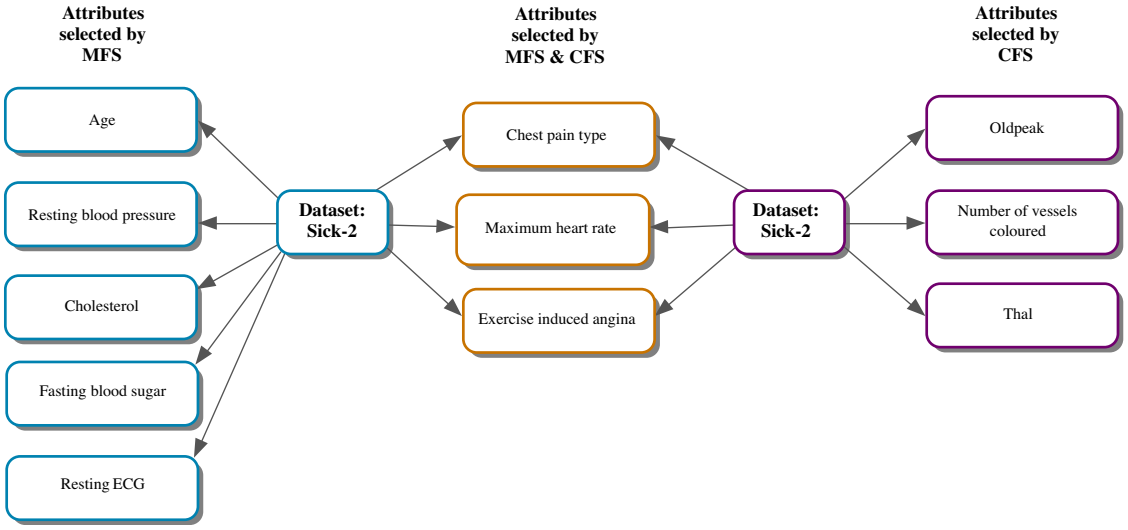

Fig. 3. Attributes selected by MFS and CFS for dataset (sick-2).

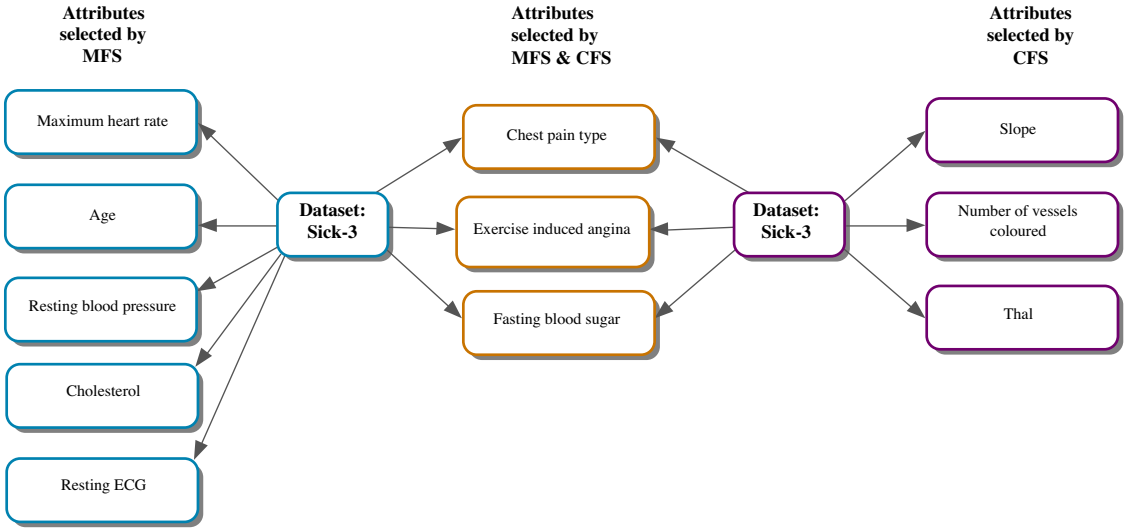

Fig. 4. Attributes selected by MFS and CFS for dataset (sick-3).

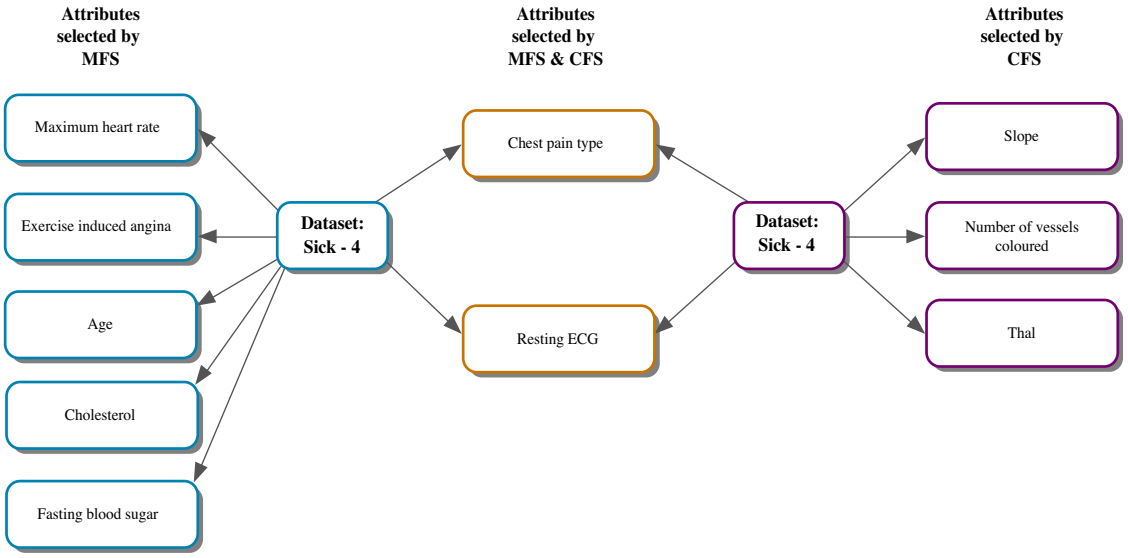

Fig. 5. Attributes selected by MFS and CFS for dataset (sick-4).

Table 4
Combining proposed feature selection (MFS) with automated feature selection (CFS).

Dataset	Algorithms	Accuracy (%)			TP			F-measure		
		MFS	CFS	CVP-10-fold	MFS	CFS	CVP-10-fold	MFS	CFS	CVP-10-fold
H-O	Naive Bayes	69.11	82.36	81.88	0.757	0.865	0.782	0.727	0.842	0.02
	SMO	77.95	77.94	75.25	0.811	0.892	0.764	0.8	0.815	0.14
	IBK	**100**	80.88	76.24	**1**	0.838	0.855	**1**	0.827	0
	AdaBoostM1	72.05	77.94	81.19	**0.784**	**0.784**	0.818	0.753	0.795	0.03
	J48	**80.88**	**80.88**	76.24	**0.838**	**0.838**	0.745	**0.827**	**0.827**	0.03
	PART	**86.77**	82.35	79.21	**0.919**	0.865	0.818	**0.883**	0.842	0.02
Sick-1	Naive Bayes	**85.29**	76.47	72.28	**0.167**	**0.167**	0.111	**0.286**	0.2	0
	SMO	**82.35**	**82.35**	82.18	**0**	**0**	0	0	0	0.16
	IBK	67.65	70.58	72.28	**0.333**	0.167	0.111	**0.267**	0.167	0.02
	AdaBoostM1	**82.35**	76.47	82.18	0	0	0	0	0	0.05
	J48	**82.35**	**82.35**	82.18	0	0	0	0	0	0.03
	PART	73.53	73.53	71.29	**0.167**	0	0.222	**0.182**	0	0.02
Sick-2	Naive Bayes	**85.29**	79.41	79.21	0	0.25	0.333	0	0.222	0.02
	SMO	**88.24**	**88.25**	88.12	0	0	0	0	0	0.16
	IBK (2)	**88.24**	82.35	79.21	**0.25**	**0.25**	0.167	**0.333**	0.25	0
	AdaBoostM1	**88.24**	82.35	86.14	0	0.25	0	0	0.25	0.03
	J48	85.29	88.23	86.14	0	0	0.167	0	0	0.02
	PART	85.29	91.18	85.15	**0.25**	**0.25**	0.167	0.286	0.4	0.02
Sick-3	Naive Bayes	82.35	100	81.19	0.25	1	0.583	0.25	1	0
	SMO	**88.24**	83.33	88.11	0	0	0	0	0	0.16
	IBK	**79.41**	75	87.13	**0.25**	0	0.333	**0.222**	0	0
	AdaBoostM1	**88.23**	83.33	86.14	0	0	0.083	0	0	0.03
	J48	**88.24**	83.33	88.12	0	0	0	0	0	0.02
	PART	**85.29**	83.33	83.17	**0.25**	0	0	**0.286**	0	0.03
Sick-4	Naive Bayes	**97.05**	94.12	92.08	0	0	0	0	0	0
	SMO	**97.05**	**97.05**	96.04	0	0	0	0	0	0.16
	IBK	**97.05**	**97.05**	95.05	0	0	0	0	0	0
	AdaBoostM1	**97.05**	**97.05**	96.04	0	0	0	0	0	0.03
	J48	**97.05**	**97.05**	96.04	0	0	0	0	0	0.02
	PART	**97.05**	**97.05**	96.04	0	0	0	0	0	0.02

sick-3 and sick-4 datasets (for sick-2, MFS was comparable to CFS in one cases and higher for three cases; for sick-3, MFS produced a higher accuracy reading than CFS in five cases, and for sick-4, MFS was higher in one case and comparable with the other five cases). Overall it is estimated that in terms of accuracy, medical based feature selection produced the better prediction performance than computerized feature selection. Table 4 also shows results in comparison to CVP 10-fold settings, used in the previous experiment (in other words, no feature selection). Results show that both MFS and CFS had improved prediction rates, in terms of accuracy, for the majority of the algorithms and for all the datasets in comparison to the results with no feature selection. The experiment showed when feature selection is applied on the UCI heart disease data, it advances the prediction efficiency, with medical knowledge based feature selection resulting in the better performance. It was also shown that using MFS with IBK resulted in better or comparable TP and F-measure results relative to the results using CFS for four datasets (H-O, sick-1, sick-2 and sick-3). Similar results were also observed using PART.

In context of the experiment it can be concluded that the results strongly suggest that using MFS improves the performance, especially in light of accuracy, of most of the classifiers for the majority

Table 5
Medical feature selection with combination of medical and computer based feature selection. (MFS stands for medical feature selection, while MFS+CFS stand for medical feature selection plus computer feature selection. Classifiers performed on the five datasets are shown, with the best performance shown in bold.)

Dataset	Algorithms	Accuracy (%)		TP		F-measure	
		MFS	MFS+CFS	MFS	MFS+CFS	MFS	MFS+CFS
H-O	Naive Bayes	69.11	**83.83**	0.757	**0.892**	0.727	**0.857**
	SMO	77.95	**83.83**	0.811	**0.919**	0.8	**0.861**
	IBK	**100**	73.53	**1**	0.784	**1**	0.763
	AdaBoostM1	72.05	**80.88**	0.784	**0.811**	0.753	**0.822**
	J48	**80.88**	73.52	**0.838**	0.784	**0.827**	0.763
	PART	**86.77**	75.00	**0.919**	0.757	**0.883**	0.767
Sick-1	Naive Bayes	**85.29**	77.94	**0.167**	0.077	**0.286**	0.118
	SMO	**82.35**	75	0	**0.077**	0	**0.105**
	IBK	67.65	**76.47**	**0.333**	0.231	0.267	**0.273**
	AdaBoostM1	**82.35**	80.88	0	0	0	0
	J48	**82.35**	80.88	0	0	0	0
	PART	**73.53**	**73.53**	**0.167**	0.077	**0.182**	0.1
Sick-2	Naive Bayes	**85.29**	**85.29**	0	***0.25***	0	**0.286**
	SMO	**88.24**	**88.24**	0	0	0	0
	IBK (2)	88.24	**94.11**	0.25	**0.5**	0.333	**0.667**
	AdaBoostM1	**88.24**	82.35	0	**0.25**	0	**0.25**
	J48	85.29	**88.23**	0	**0.5**	0	**0.5**
	PART	**85.29**	**85.29**	**0.25**	**0.25**	**0.286**	**0.286**
Sick-3	Naive Bayes	82.35	**85.30**	0.25	**0.75**	0.25	**0.545**
	SMO	**88.24**	**88.23**	0	0	0	0
	IBK	79.41	**85.29**	**0.25**	0	**0.222**	0
	AdaBoostM1	**88.23**	85.30	0	0	0	0
	J48	**88.24**	**88.23**	0	0	0	0
	PART	85.29	**88.23**	0.25	**0.5**	0.286	**0.5**
Sick-4	Naive Bayes	**97.05**	91.17	0	0	0	0
	SMO	**97.05**	**97.05**	0	0	0	0
	IBK	**97.05**	**97.05**	0	0	0	0
	AdaBoostM1	**97.05**	**97.05**	0	0	0	0
	J48	**97.05**	**97.05**	0	0	0	0
	PART	**97.05**	**97.05**	0	0	0	0

of the datasets. Therefore, the method shows promise in case of measured performance. In the next section, MFS is investigated further in combination with CFS.

8. Extension to MFS

The previous section showed that MFS is a promising feature selection method in heart disease diagnosis and that CFS often disregards features that are medically important. This section details an extension to the MFS. Features selected by CFS were combined with those selected by MFS and the results were compared to using MFS alone. The objective was to see if this combination of MFS and CFS improved prediction results over MFS. Results are shown in Table 5. The experimental results for combining MFS and CFS are shown using the symbol MFS+CFS.

The results showed that in terms of accuracy, MFS improved the performance for 10 cases (for H-O: IBK, J48 and PART-86.77; for sick-1: Naive Bayes, SMO, AdaBoostM1 and J48; for sick-2: AdaBoostM1; for sick-3: AdaBoostM1; for sick-4: Naive Bayes) and 11 comparable accuracy to MFS+CFS among all the datasets. MFS+CFS indicated improved accuracy for nine aspects within H-O: Naive Bayes, SMO and AdaboostM1; for sick-1: IBK; for sick-2: IBK and J48 and for sick-3: Naive Bayes, IBK and PART. Although the performance of both methods appeared comparable, in terms of accuracy MFS showed marginally higher results. But in terms of TP and F-measure, MFS+CFS resulted in a higher performance than using MFS alone for the majority of the cases and particularly for the Naive Bayes classifier (better than MFS for three datasets). On the other hand, for MFS+CFS, SMO performed better in terms of accuracy for one of the dataset (H-O), and comparable for three of the datasets (sick-2, sick-3, and sick-4). SMO also performed better using MFS+CFS in terms of accuracy than using MFS for one dataset (H-O) and performed comparably to MFS for three datasets (sick-2, sick-3, sick-4). Similar results were also observed for the IBK algorithm. SMO, however, further showed some improvement in terms of TP and F-measure using MFS+CFS over using MFS alone. Overall, the performance indicated that MFS feature selection appeared to be a favorable strategy for heart disease data. But the combination of MFS and CFS has promise for some of the classifiers (particularly Naive Bayes, IBK and SMO). Therefore, the proposed MFS and MFS+CFS are promising techniques for use in heart disease diagnostics.

9. Conclusion

Early detection of heart disease is essential to save lives. Understanding the usefulness of data mining for assisting in the diagnosis of heart disease is so important. This paper has provided details on the comparison of classifiers for the detection of heart disease. It was observed that SMO (Support Vector Machine) has shown potential classification algorithm in this area, particularly when considering total accuracy as a performance measure. The paper also presented outcomes from using automated feature selection and a medical knowledge based motivated feature selection process (MFS). The results of experiment demonstrated that the use of MFS noticeably improved the performance especially in terms of accuracy, for most of the classifiers for the majority of the datasets. This indicates that the method has promise. MFS combined with the computerized feature selection process (CFS) was also experimented and encouraging results were seen for some of the classifiers particularly NaiveBayes, IBK and SMO. In summary, the MFS and MFS+CFS are promising techniques for use in heart disease diagnostics.

References

Abdel-aal, R. (2005). Improved classification of medical data using abductive network committees trained on different feature subsets. *Computer Methods and Programs in Biomedicine, 80*, 141–153.

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

Aha, D., & Kibler, D. (1988). Instance-based prediction of heart-disease presence with the Cleveland database. Technical Report, University of California, Irvine, Department of Information and Computer Science, Number ICS-TR-88-07.

Avci, E. (2009). A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier. *Expert Systems with Applications, 36*, 10618–10626.

Baek, S., Tsai, C., & Chen, J. (2009). Development of biomarker classifiers from high-dimensional data. *Briefings in bioinformatics, 10*, 537–546.

Baliga, R. R., & Eagle, K. A. (2008). *Practical cardiology: Evaluation and treatment of common cardiovascular*. Lippincott Williams & Wilkins.

Boors, E., Hammer, P., Ibaraki, T., Kogan, A., Mayoraz, E., & Muchnik, I. B. (2000). An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering, 12*, 292–306.

Chang, C. (2010). Recognition of atrial fibrillation and congestive heart failure based on heart rate variability. Masters thesis, Graduate Institute of Electrical Engineering, China.

Cheng, T., Wei, C., & Tseng, V. (2006). Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. In *19th IEEE symposium on computer-based medical systems (CBMS'06)* (pp. 165–170).

Chen, S., Zhou, S., Zhang, J., Yin, F., Marks, L., & Das, S. (2007). A neural network model to predict lung radiation-induced pneumonitis. *Medical physics, 34*, 3808–3814.

Dash, M. O. Y. H. (2008). Efficient cross validation over skewed noisy data. *IEEE International Conference on Systems, Man and Cybernatics (SMC)*, 749–756.

Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications, 36*, 7675–7680.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology, 64*, 304–310.

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).

Diagnosis, E. (2010). Coronary disease or heart attack from expert system: Chest pain [Online]. Available: <http://www.easydiagnosis.com/cgi-bin/expert/explain2.cgi?mod=Chest+Pain&ask=ddisease2&title=Coronary+Disease+or+Heart+Attack&showmod=yes> [Accessed 12th December 2010].

Doyle, O., Temko, A., Marnane, W., Lightbody, G., & Boylan, G. (2010). Heart rate based automatic seizure detection in the newborn. *Medical Engineering and Physics, 32*, 829–839.

Edmonds, B. H. (2005). *Using localised 'Gossip' to structure distributed learning, centre for policy modelling. Proceedings of the joint symposium on socially inspired computing engineering with social metaphors (AISB)*. Hatfield, UK: University of Hertfordshire. 127–134.

El-hanjouri, M., Alkhaldi, W., Hamdy, N., & Alim, O. A. (2002). Heart diseases diagnosis using HMM. In *11th mediterrane electrotechnical conference*, MELECON, Cairo, Egypt (pp. 489–492).

Fan, Y., & Chaovalitwongse, W. (2010). Optimizing feature selection to improve medical diagnosis. *Annals of Operations Research, 174*, 169–183.

Fountoulaki, A., Karacapilidis, N., & Manatakis, M. (2010). Using decision trees for the semi-automatic development of medical data patterns: A computer-supported framework. *Web-Based Applications in Healthcare and Biomedicine*, 229–242.

Freedc. 2010. Asymptomatic [Online]. Available: <http://www.thefreedictionary.com/asymptomatic> [Accessed 12th December, 2010].

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).

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence, 40*, 11–61.

Hanbay, D. (2009). An expert system based on least square support vector machines for diagnosis of the valvular heart disease. *Expert Systems with Applications, 36*, 4232–4238.

Kaul, U. (2010). What is typical and atypical angina? [Online]. Available: <http://doctor.ndtv.com/faq/ndtv/fid/2907/What_is_typical_and_atypical_angina.html> [Accessed 12th February, 2010].

Khan, M. I. G. (2005). *Heart disease diagnosis and therapy: A practical approach*. Humana Press.

Kim, B.-H., Lee, S.-H., Cho, D.-U., & Oh, S.-Y. (2008). A proposal of heart diseases diagnosis method using analysis of face color. In *International conference on advanced language processing and web information technology, ALPIT* (pp. 220–225).

Kumar, K., & Shelokar, P. (2008). An SVM method using evolutionary information for the identification of allergenic proteins. *Bioinformation, 2*, 253–256.

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.

Mei, H., Ma, D., Ashley-koch, A., & Martin, E. (2005). Extension of multifactor dimensionality reduction for identifying multilocus effects in the GAW14 simulated data. *BMC genetics, 6*, S145.

Mengel, M. B., & Schwiebert, L. P. (2005). *Family medicine: Ambulatory care & prevention*. McGraw-Hill Professional.

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).

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In *International conference on computer systems and applications, AICCSA*. IEEE/ACS, Doha IEEE (pp. 108–115).

Pickett, J. P. (2000). *The American heritage dictionary of the english language*. Houghton Mifflin.

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.

Polat, K., & Güne, S. (2007). An improved approach to medical data sets classification: Artificial immune recognition system with fuzzy resource allocation mechanism. *Expert Systems, 24*, 252–270.

Qazi, M., Fung, G., Krishnan, S., Rosales, R., Steck, H., Rao, B., et al. (2007). Automated heart wall motion abnormality detection from ultrasound images using Bayesian networks. In *International joint conference on artificial intelligence* (pp. 519–525).

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

Skalak, D. B. (1997). Prototype selection for composite nearest neighbor classifiers. Department of Computer Science, University of Massachusetts-Amherst, PhD thesis.

Smith, J. S. (2010). Screening for high-risk cardiovascular disease: A challenge for the guidelines. *Archives of Internal Medicine, 170*, 40–42.

Society, M. M. (1945). *The New England journal of medicine*. 232: Massachusetts Medical Society, New England Surgical Society, HighWire Press new.

Soman, K. P., Shyam, D. M., & Madhavdas, P. (2003). Efficient classification and analysis of ischemic heart disease using proximal support vector machines based decision trees, TENCON. In *Conference on convergent technologies for Asia-Pacific region* (Vol. 1, pp. 214–217).

Uci. 2009. Heart disease dataset [Online]. Available: <http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/cleve.mod> [Accessed 5th March, 2009].

Uci. 2010. Cleveland heart disease data details [Online]. Available: <http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names> [Accessed 8th February 2010].

Wang, Y., & Ma, L. (2009). Feature selection for medical dataset using rough set theory. In *Proceedings of the 3rd WSEAS international conference on computer engineering and applications (CEA)* (pp. 68–72). Stevens Point, Wisconsin: USA World Scientific and Engineering Academy and Society (WSEAS).

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

Xing, Y., Wang, J., Zhao, Z., & Gao, A. (2007). Combination data mining methods with new medical data to predicting outcome of coronary heart disease (PDF). In *International conference on convergence information technology (ICCIT 2007)*.

Zhao, H., Chen, J., Hou, N., Zheng, C., & Wang, W. (2010). Identifying metabolite biomarkers in unstable angina in-patients by feature selection based data mining methods. In *Second international conference on computer modeling and simulation* (pp. 438–442). Sanya, China: IEEE.

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.

Zheng, J., Jiang, Y., & Yan, H. (2006). Committee machines with ensembles of multilayer perceptron for the support of diagnosis of heart diseases. In *Proceedings of the international conference on, communications, circuits and systems* (Vol. 3, pp. 2046–2050).