Journal of Computer Science and Cybernetics, V.34, N.1 (2018), 33–47
DOI: 10.15625/1813-9663/34/1/12665

AUTOMATIC HEART DISEASE PREDICTION USING FEATURE SELECTION AND DATA MINING TECHNIQUE

LE MINH HUNG[1,a], TRAN DINH TOAN[1], TRAN VAN LANG[2]

[1]*Information Technology Faculty, Ho Chi Minh City University of Food Industry*
[2]*Institute of Applied Mechanics and Informatics, VAST*
[a]*hunglm@cntp.edu.vn*

Abstract. This paper presents an automatic Heart Disease (HD) prediction method based on feature selection with data mining techniques using the provided symptoms and clinical information assigned in the patients dataset. Data mining which allows the extraction of hidden knowledges from the data and explores the relationship between attributes, is the promising technique for HD prediction. HD symptoms can be effectively learned by the computer to classify HD into different classes. However, the information provided may include redundant and interrelated symptoms. The use of such information may degrade the classification performance. Feature selection is an effective way to remove such noisy information meanwhile improving the learning accuracy and facilitating a better understanding for learning model. In our method, HD attributes are weighted and re-ordered based on their rank and weights assigned by Infinite Latent Feature Selection (ILFS) method. A soft margin linear Support Vector Machine (SVM) is applied to classify a subset of selected attributes into different HD classes. The experiment is performed using UCI Machine Learning Repository Heart Disease public dataset. Experimental results demonstrated the effectiveness of the proposed method for precise HD prediction making, our method gained the best performance with an accuracy of 90.65% and an AUC of 0.96 for distinguishing 'No presence' HD with 'Presence' HD.

Keywords. Data mining, Heart Disease Prediction, Feature Selection, Classification.

1. INTRODUCTION

Heart disease (HD) is one of the top leading causes of death accounting for 17.7 million deaths each year, 31% of all global deaths, as reported by World Health Organization 2017. Patients unhealthy habits such as tobacco use, unhealthy diet, physical inactivity and alcohol usage are the main reasons leading to many types of HD. Several clinical information and symptoms are found to be related to HD including age, blood pressure, total cholesterol, diabetes, hyper tension [1]. HD dataset basically consists of the above-mentioned information and attributes which summarized and collected from the patients. With the increasing of the huge amounts of dataset made available in recent years, the diagnosis of HD can be automatically performed using traditional statistical methods to predict the potential of having HD on each patient. Working with HD database can be considered as a real-life application and learning such attributes helps clinicians in identifying the main risk factors associated with HD. However, with a large number of attributes, it is challenging to identify

which attributes are the most significant risk factors for HD prediction by just only based on conventional statistical methods.

To tackle this problem, there have been numerous dedicated approaches based on data mining techniques proposed in recent years to help healthcare professionals in the diagnosis of HD. HD prediction systems based on data mining techniques could assist doctors in giving accurately HD prediction making based on the clinical information data of patients. Data mining techniques which refers to mining the information, allow the extraction of hidden knowledge and establishe the relationships between attributes inside the data, is the promising techniques for HD prediction [2, 3, 4]. Such invention could assist doctors in better health policy-making, prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths. Specifically, Deepika et al. proposed association rule for classification of Heart-attack patients [5]. K. Srinivas et al. presented data mining techniques in Healthcare and Prediction of Heart Attacks based on Naive Bayes algorithm, K-NN, Decision Tree, wherein Decision Tree achieved the best performance among the methods [6]. Similarly, several classification algorithms including Naive Bayes, Decision Tree, and Neural Network were compared in [7] for the prediction of stroke diseases. The experimental results showed that the Neural Network performed much better than the other two algorithms. Jabbar et al. proposed association rule mining for heart attack prediction based on the sequence number and clustering, in which the patterns are extracted from the database with significant weights calculation [8]. Shouman et al. combined k-means clustering with decision tree method to predict the HD on a subset of 13 input attributes [9]. This study suggested that integrating k-means clustering and decision tree could achieve a higher accuracy than other traditional methods in the diagnosis of HD patients. Dangare et al. proposed an improved Study of Heart Disease Prediction System using Data Mining Classification Techniques [10]. Their purpose was to build an Intelligent Heart Disease Prediction System that gives diagnosis of HD by using historical heart database such as sex, blood pressure, cholesterol, obesity and smoking, etc. Neural networks were adopted for the classification of 14 attributes by considering the single and multilayer neural network models in [11]. Olatubosun et al. [12] proposed to use Artificial Neural Network with back propagation procedure for the diagnosis of Cerebrovascular disease. M. Anbarasi et al. proposed Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm [13]. Classification techniques such as Naive Bayes, Decision Tree and Classification were adopted, in which Naive Bayes achieved the highest performance across the methods. Patel et al. [14] proposed to use the reduced number of attributes using tree classification function techniques in data mining including Naive Bayes, Decision Tree and Classification by Clustering, in which Decision Tree gained the best performance among the methods.

For feature selection, Singh et al. [15] proposed to use Genetic feature selection method combined with Naive Bayes method for HD prediction. Takci searched for the best machine learning method and feature selection method for heart attacks prediction, in which SVM with linear kernel in combination with Relief-Based Feature achieved the best performance [16]. However, this study used a small number of dataset with 270 instances and a limited number of HD attributes (13 attributes). Similarly, Suganya et al. proposed a novel feature selection method for Cardiac diseases prediction on the selected 13 attributes with a total of 303 instances of patients dataset [17]. Mirmozaffari et al. applied clustering methods integrated in WEKA data mining tool on a patients dataset with 8 attributes and a total

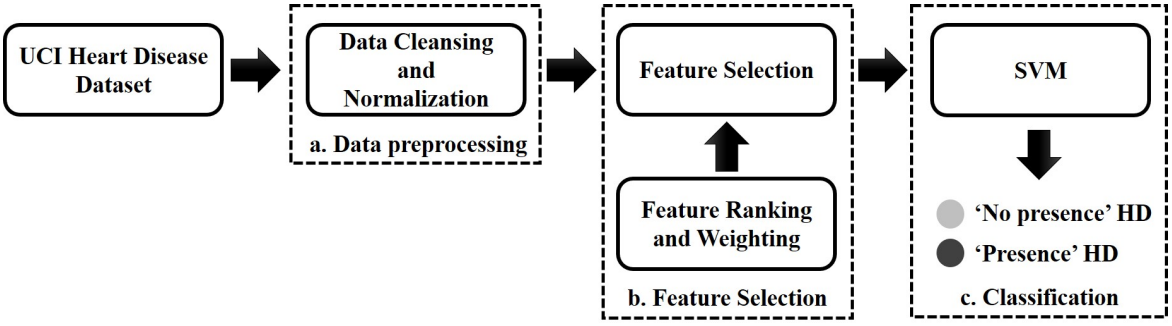

Figure 1. Our 3-step proposed feature selection for data mining in HD diagnosis. (a) Step 1: Data preparation. (b) Step 2: Feature Selection. (c) Step 3: Classification

of 209 instances for heart disease prediction [18]. Uma et al. applied several classification algorithms (e.g. SVM, Bagging, Naive Bayes, Regression, J48) and feature selection methods (e.g. CfsSubsetEval, Information Gain, Gain Ratio and Wrapper method) on a subset of 18 attributes of HD on a dataset with a total of 689 instances [19]. They proved that SVM achieved the best performance among the classifiers and most of the adopted feature selection methods achieved nearly identical accuracy.

Despite various approaches have been proposed for HD prediction, most of the recent feature selection methods were designed on a small subset of attributes with 14 attributes or 6 attributes. There is still a lack of effective methods based on feature selection and data mining techniques to study the significant risk factors associated with HD on the fully provided attributes. There might be existing other hidden factors or attributes that play an important role on making HD prediction, which has not yet been comprehensively explored in previous studies. In this work, we proposed a method to efficiently and effectively predict different classes of HD based on feature selection and data mining technique. The HD diagnosis prediction task in this study is distinguishing between 'No presence' HD (labeled as 0 in the dataset) and 'Presence' HD (labeled as 1, 2, 3, 4 in the dataset). Our method consists of three main steps which are: Step 1: Data Preparation; Step 2: Feature Selection; and Step 3: Classification. Specifically, the unnecessary and noisy attributes are first manually removed in the step 1. Then feature selection based ILFS described in [20] is adopted to select the most significant attributes based on the extracted weights and rank. These selected useful attributes could drastically affect the performance of the prediction diagnosis system. A soft-margin linear kernel Support Vector Machine (SVM) is finally applied to classify the subset of selected attributes into two classes of 'No presence' and 'Presence' HD. Our contributions can be highlighted as follows:

- We performed feature selection with data mining methods on the fully provided attributes of HD with a larger number of instances (699 instances), which is different from previous studies which mainly based on a given subset of attributes (e.g. 13 attributes or 6 attributes) and a limited number of patient dataset.

- We applied ILFS feature selection method based on [20] to select the most discriminative and meaningful attributes used for the HD prediction making. We found that

by using only an approximately half of the given HD attributes selected by ILFS, the prediction performance is competitive compared with using the fully given HD attributes. This demonstrated that the HD dataset contains more redundant attributes which play less important roles for prediction making.

- We found that different feature selection methods select different attributes for HD prediction and the performance varies quite differently. The choice of feature selection methods may depend on the availability of the given number of the attributes to achieve a desirable performance.
- The proposed method can be feasibly applied and integrated in many healthcare diagnosis systems for disease prediction making as well as real-life applications. The source code of our method will be made available with the publication of this paper.

The rest of the paper is organized as follows: Sec. 2 describes in details our 3-step method for HD prediction. Sec. 3 summarizes the results from our method. Sec. 4 is the discussion of our paper.

2. METHODOLOGY

As illustrated in Fig. 1, our proposed method, which demonstrates an excellent agreement with the manually assigned labels, consists of 3 main steps. Firstly, irrelevant attributes and noisy information are manually removed from the original raw dataset and only the most meaningful attributes are preserved. ILFS [20] for feature ranking and feature selection is utilized in step 2 to select a subset of discriminative attributes, i.e. the most significant risk factors associated with HD. A supervised SVM with soft-margin linear kernel is finally used to classify the selected attributes into different classes.

2.1. Data preparation

Irrelevant attributes are firstly manually removed from the original dataset. As a result, 58 attributes are preserved from the provided original 75 attributes in each instance as described in details in Table 1. To reduce the inhomogeneity in each attribute among the patients, the numeric-valued numbers assigned in each attribute is normalized by z-score method. The dataset is organized in the form of a matrix with the size of $N \times M$, where N is the number of patients and M is the number of attributes ($N = 699, M = 58$ in this study). After preprocessing, 80% of the dataset is selected for training and the remaining 20% of the dataset is used for testing.

2.2. Feature selection

It is worth noticing that most of the real-life data contains more information than it is needed to build a model, or the wrong kind of information. Noisy or redundant information makes it more challenging to extract the most meaningful information. Feature selection which refers to the process of reducing the inputs for processing and analysis, or finding the most meaningful subset of information, is effective for the prediction performance. Feature selection does not only improve the quality of the model but also makes the process

of modeling more efficient. The most highlighted techniques proposed recently can be referred to is Recursive Feature Elimination Support Vector Machine (RFE-SVM) [22] which successfully applied in the application of prostate cancer diagnosis to reduce the dimension of hand-crafted features extracted from the lesion region of interest and achieved a very high accuracy compared with using the fully dimension of data attributes [23]. However, in the work [20] which provides a more comprehensive overview of feature selection techniques, ILFS achieved the best performance among the 14 popular feature selection methods.

Inspired by [20], ILFS was adopted to select the most discriminative attributes of the feature vectors used for HD prediction in our paper. ILFS allows the selection of a subset of features expected to be most likely to discriminate between classes of HD. The HD attributes weights and rank are automatically assigned based on ILFS method. Weights are assigned by a Graph-weighting which is basically based on the undirected fully connected graph and automatically learnt based on a learning framework on the probabilistic latent semantic analysis (PLSA) [24]. Expectation Maximization algorithm is adopted to estimate the parameters. The ranking step is built based on Infinite Feature Selection [25] filter algorithm in an unsupervised manner, followed with the cross-validation strategy for selecting the best subset of features. Specifically, suppose $X = \left\{X_1, X_2, , X_n\right\}$ is a set of given training features, m as the number of samples, $m \times 1$ vector X_i is the distribution of the values assumed by i^{th} feature. Weights are associated with the undirected graph nodes

$$a_{ij} = \varphi(X_i, X_j) \tag{1}$$

where a_{ij} is the node corresponding to features and edges model relationship between any pairs of nodes, $\varphi(X_i, X_j)$ is considered to be a real-valued function learned by the probability of each co-occurrence in X_i, X_j as a mixture of an independent multinomial distributions. Each weight represents the likelihood that features X_i and X_j are good candidates. For further details about the ILFS algorithm, interested readers can refer to [20]. We implemented ILFS based on the MATLAB code provided in Feature Selection Library (FSLib 2017) [26].

2.3. Classification

A linear supervised SVM classifier is applied to map the selected attributes into 2 classes of 'No presence' and 'Presence' HD. The basic idea of the SVM is to construct a hyperplane to separate and maximize the margin of the positive and negative classes with the largest margin. Suppose $\left\{(x_i, y_i)\right\}_{i=1}^{N}$ is a set of training samples which contain the most discriminant attributes selected by ILFS, (x_i, y_i) is the input feature for the i^{th} instance and its the corresponding target output, respectively. The decision boundary separates the instances by the equation form

$$w^T x_i + b \geq 0 \quad \text{for} \quad y_i = +1 \quad \text{(positive class)}, \tag{2}$$

$$w^T x_i + b < 0 \quad \text{for} \quad y_i = -1 \quad \text{(negative class)}, \tag{3}$$

where w is an adjustable weight vector, x is an input vector, and b is a bias. Assume that the features selected by ILFS are linear separable, the optimization problem of SVM to maximize

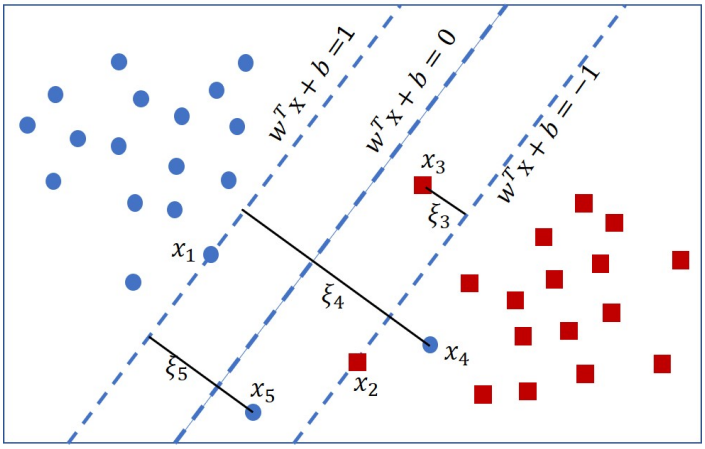

Figure 2. SVM with soft margin kernel with different cases of slack variables.

the margin can be defined as

$$(w, b) = \arg\min_{w,b} \frac{1}{2}\|w\|_2^2 \text{ s.t } y_i(w^T.x_i + b) \geq 1, \ \forall i = 1, 2, ..., N. \tag{4}$$

The normal SVM normally works with the linear separable features. However, in some cases when there exist noises which belong to one class but appear closely to another class, even if the two classes are linear separable, SVM in this scenario will construct a hyperplane with a very small margin, which is very sensitive to noise. If the algorithm sacrifices these noises, SVM could generate a better hyperplane with a better margin to best separate the two classes. Another scenario is when the two classes are near linear separable, in which there exist a small number of instances appeared unproperly, the optimization algorithm of SVM margin is infeasible. Similarly, if the algorithm ignores those instances, SVM could also generate a better margin that could mostly separate the two classes. This technique called SVM with soft margin. The formulization of the SVM optimization problem can be re-written as follows

$$(w, b, \xi) = \arg\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} \xi_i \text{ s.t } 1 - \xi_i - y_i(w^T.x_i + b) \geq 1, \tag{5}$$

$$\forall i = 1, 2, ..., N, \ \xi_i \geq 0, \ C > 0,$$

where C is the regularization term used to avoid overfitting, $\xi = [\xi_1, \xi_2, ..., \xi_N]$ is a set of slack variables. As shown in Fig. 2, for the variables which are located in the safety margin, then $\xi_i = 0$ (e.g. x_1, x_2). For the variables which are not located in the safety margin, but still in the right side of their class, then $0 < \xi_i < 1$ (e.g. x_3). For the variables which are located in the wrong side of their class, then $\xi_i > 1$ (e.g. x_4, x_5).

Table 1. Description of 58 HD attributes used for HD prediction

No.	Attribute	No.	Attribute
1	age	30	tpeakbpd: peak exercise blood pressure
2	sex	31	trestbpd: resting blood pressure
3	painloc: chest pain location	32	exang: exercise induced angina (1 = yes; 0 = no)
4	painexer (1 = provoked by exertion; 0 = otherwise)	33	xhypo: (1 = yes; 0 = no)
5	relrest (1 = relieved after rest; 0 = otherwise)	34	oldpeak = ST depression induced by exercise relative to rest
6	cp: chest pain type	35	slope: the slope of the peak exercise ST segment
7	trestbps: resting blood pressure	36	rldv5: height at rest
8	htn	37	rldv5e: height at peak exercise
9	chol: serum cholesterol in mg/dl	38	ca: number of major vessels (0-3) colored by fluoroscopy
10	cigs (cigarettes per day)	39	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
11	years (number of years as a smoker)	40	thalsev
12	fbs: (fasting blood sugar > 120 mg/dl)	41	thalpul
13	famhist: family history of coronary artery disease	42	cmo: month of cardiac
14	restecg: resting electrocardiographic results	43	cday: day of cardiac
15	20 ekgmo (month of exercise ECG reading)	44	cyr: year of cardiac
16	ekgday(day of exercise ECG reading)	45	lmt
17	ekgyr (year of exercise ECG reading)	46	ladprox
18	dig (digitalis used furring exercise ECG)	47	laddist
19	24 prop (Beta blocker used during exercise ECG)	48	diag
20	nitr (nitrates used during exercise ECG)	49	cxmain
21	pro (calcium channel blocker used during exercise ECG)	50	ramus
22	diuretic (diuretic used during exercise ECG)	51	om1
23	proto: exercise protocol	52	om2
24	thaldur: duration of exercise test in minutes	53	rcaprox
25	thaltime: time when ST measure depression was noted	54	rcadist
26	met: mets achieved	55	lvx3
27	thalach: maximum heart rate achieved	56	lvx4
28	thalrest: resting heart rate	57	lvf
29	tpeakbps: peak exercise blood pressure	58	cathef

3. EXPERIMENTAL RESULTS

3.1. Datasets

The HD database used in our study is the public dataset collected from UCI Machine Learning Repository [21]. This directory consists of 4 HD datasets collected from 4 different hospitals, which include

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

We select 3 datasets with the total number of 699 instances including the Cleveland dataset (282 instances), Hungarian dataset (294 instances) and the Switzerland dataset (123 instances) dataset. The instances in the original dataset are labeled into 5 different classes in which class 0 indicates 'No presence' HD and class 1 to class 4 indicate the risk levels of HD, denoted as 'Presence' HD. Finally, a total number of instances of the two classes 'No presence' and 'Presence' HD are 353 and 346, respectively. The UCI Heart Disease database has been examined by professional clinicians and widely used in many previous data mining-based approaches for HD prediction. 76 raw attributes presented as numeric-valued numbers in each row are the collection of different diagnosis attributes and medical information collected from each patient. Unlike most of the recent studies which just only investigate a subset of 14 attributes or 6 attributes from this database, our study fully explores most of the provided information in the original dataset (except for the attribute with missing values).

3.2. Experimental designs

In this section, we conducted 2 experiments to investigate the performance of several classification and feature selection methods. In the experiment 1, different classification methods are performed to select the most reliable method for HD prediction. The characteristic of the HD dataset is also analyzed in this experiment. The selected classification method is then utilized in the experiment 2 to classify the selected attributes of HD into two classes. To avoid overfitting, the validation of all the methods are performed using the hold-out strategy, where the dataset is randomly split into 2 independent parts for training (80%) and testing (20%). We selected the hold-out strategy instead of k-fold cross validation since the hold-out strategy avoids the overlap between training set and testing set, which provides a more accurate estimate for the generalization performance of the algorithm. With k-fold cross validation strategy, the feature selection and classification have to be performed independently k times yielding k feature rankings and k models, respectively. With the limited number of the given dataset, the ranking of the features given by the same feature selection algorithm may be slightly inconsistent for each running time, which is not feasible for the testing.

Experiment 1: Classifiers comparisons

To evaluate the effectiveness of our proposed method, we compare our method with 4 classification methods including Non-linear SVM (Polynomial kernel, Gaussian kernel, and Sigmoid kernel), Nave Bayes and Logistic regression classifier. Nave Bayes and Logistic regression classifiers are performed using the WEKA data mining tool, which is an open source software issued under the GNU General Public License and is a very popular software for solving data mining problems. WEKA also includes a collection of machine learning algorithms for data mining tasks and has been adopted in many data mining applications due to its simplicity and friendly user interface. SVM algorithms with linear and non-linear kernels are implemented using Matlab (Release 2017a, Natick MA) on a PC running on a single Intel core i7 CPU, Windows 10 OS.

Experiment 2: Feature selection methods comparisons

In this experiment, several feature selection methods are selected for the performance comparisons including:

- Principle Component Analysis denoted as PCA [27]: is one of the most important unsupervised statistical procedure in machine learning for dimensionality reduction or feature selection. The goal of PCA is to find the best representation of the data by projecting them onto a lower dimension space called principal components (PCs), in which the first PC has the largest variance and so on. Ziasabounchi et al. successfully applied PCA together with k-means clustering in the application of heart disease prediction [28].

- Sort features according to pairwise correlations which is denoted as CFS [29]: CFS is a simple filter algorithm which ranks the features based on the correlation with the class labels and select the most informative features subset which highly correlated used for classification. CFS is based on the assumption that good features are highly correlated with the classification and not correlated to each other.

- Feature Selection and Kernel Learning for Local Learning-Based Clustering denoted as LLCFS [30]: is an unsupervised clustering feature selection method which considers the relevance of each feature for clustering based on a built-in structure learning procedure to iteratively update the Laplacian graph. Feature weight learning process is performed using the constructed k-nearest neighbor graph built on the weighted feature space.

3.3. Results

Accuracy, sensitivity and specificity are used as the evaluation metrics to evaluate the classification performance of our HD diagnosis prediction system. Area under the curve (AUC) of receiver operating characteristics (ROC) is also provided for the binary classification. The classification accuracy, sensitivity and specificity are defined as follows

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{7}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{8}$$

where TP, FP, TN, FN are true positive, false positive, true negative and false negative, respectively. In this study, we consider the best performance in term of *Accuracy* and AUC.

Experiment 1

Table 2. Results of HD prediction using different classifiers

Methods	**Accuracy (%)**	**AUC**	**Sensitivity**	**Specificity**
Linear SVM	89.93	0.96	0.87	0.93
Non-linear SVM (Gaussian)	49.64	0.66	0.00	1.00
Non-linear SVM (Polynomial)	83.45	0.92	0.85	0.81
Non-linear SVM (Sigmoid)	49.64	0.41	0.00	1.00
Nave Bayes	77.70	0.86	0.64	0.91
Logistic regression Classifier	85.61	0.91	0.81	0.90

We performed 3 classification methods on the selected 58 attributes from the original dataset. Table 2 summarizes the results comparison among the methods, in which SVM with linear kernel generates the best performance with an accuracy of 89.21% and an AUC of 0.95. Logistic regression classifier achieved a competitive result with an accuracy of 85.61% and an AUC of 0.91 followed with Nave Bayes with an accuracy of 76.98% and an AUC of 0.86. SVM with Gaussian and Sigmoid kernels fail to predict the two classes of HD. Although, the performance of using SVM with Polynomial kernel could generate a high result with an accuracy of 83.45% and AUC of 0.92, the performance achieved is still lower than using linear kernel. It is maybe because of the overfitting problem caused when the hyperplane of SVM is too fit to the data, which is too sensitive to the data. The results demonstrate the effectiveness of soft-margin linear SVM in the classification task of HD. The results also show that the attributes of HD dataset can be considered as linear-separable and a linear SVM with soft margin is feasible for making precise prediction for HD. According to the results, we select SVM as a classifier to perform in the next experiment where SVM is used to classify a subset of the selected attributes extracted from feature selection methods.

Experiment 2

In order to intuitively visualize the effect of the selected attributes on the HD prediction, we plotted the accuracy and AUC curves according to the number of attributes selected by different feature selection methods, as shown in Fig. 3. Overall, the performance of all the feature selection algorithms increase when the numbers of attributes increase. According to Fig. 3, it can be observed that for the number of the selected attributes ranging from 1 to 31, PCA yields a better performance compared with other methods. However, the performance of PCA downgrades when the number of selected attributes increases until it reaches the best performance using 58 PCs with an accuracy of 89.93% and an AUC of 0.96.

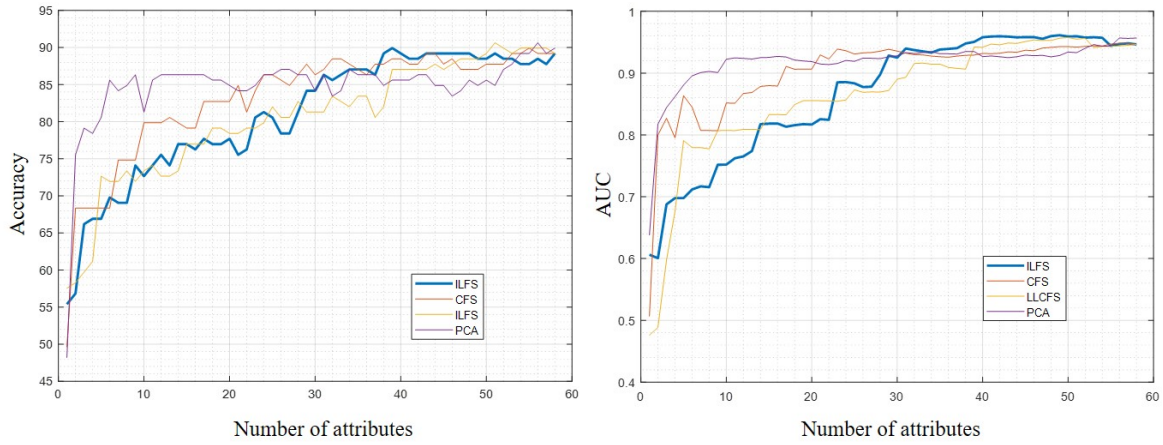

Figure 3. Accuracy and AUC of different feature selection methods with different number of selected attributes for HD prediction

Table 3. The best performances of HD prediction using different methods

Methods	**Number of attributes**	**AUC**	**Accuracy (%)**	**Sensitivity**	**Specificity**
ILFS	39	0.96	90.65	0.91	0.90
CFS	55	0.95	89.93	0.91	0.89
LLCFS	57	0.95	89.93	0.93	0.90
PCA	58	0.96	89.93	0.92	0.88

For a subset with the number of selected attributes larger than 31, ILFS performed the best and maintained stable in term of AUC, which reflects the effectiveness of ILFS in selecting and re-ordering the attributes to best optimize the classification performance. The classification accuracy achieved by CFS is competitive with ILFS when using a subset of over 31 attributes and the performance of LLCFS increases when the number of attributes increases. Table 3 summarizes the best performance achieved from the feature selection methods. It can be observed that ILFS achieved the best performance by only using 39 attributes of HD, and the performance is slightly higher than using the fully 58 attributes in term of accuracy. Although the effect of ILFS is not negligible for the incremental of the performance, ILFS only uses 39 selected attributes to achieve the best performance discarding the remaining 19 attributes.

4. DISCUSSION

In this study, we have conducted 2 experiments to investigate the performance of HD prediction using different classification and feature selection methods. Although the HD dataset can be considered as linear separable, a hard-margin SVM hardly separates the two classes. Soft-margin kernel SVM is selected as the classifier to compare the effectiveness of 4

different feature selection methods including ILFS, CFS, LLCFS and PCA. Our experiments results show that PCA could generate a competitive result when the number of PCs used is less than 31 while CFS and LLCFS perform well with over 31 attributes. ILFS generates the best performance and maintains stable when the number of attributes used is over 31.

Although ILFS could effectively select and combine a set of attributes to best optimize the classification performance, in which the best performance is recorded when using the 39 attributes selected by ILFS. However, it can be observed that using a subset with the number of attributes ranging from 31 to 39 is still feasible since the performance in this range of attributes is still reliable with the average of accuracy is approximately 88% and average of AUC is approximately 0.94. This is interesting when the doctors can only work with an approximately half of the given number of attributes but still achieve competitive results compared with using fully given attributes. This helps to reduce the workloads and time for doctors and to avoid other unnecessary clinical measurements for patients.

It can be observed that the performance of CFS and LLCFS achieved the best performance when using a large number of attributes, e.g. 55 for CFS and 57 for LLCFS. Compared with LLCFS, the performance when using CFS is more stable when the number of attributes is larger than 31 with an average of accuracy approximately 86% and an average of AUC approximately 0.92. However, CFS performed much better than ILFS when the number of selected attributes ranging from 7 to 30. For instance, the performance by only using 13 attributes selected by CFS can achieve an accuracy of 80.56% and an AUC of 0.87 while the performance when using ILFS only achieve an accuracy of 74.10% and an AUC of 0.77. The results demonstrated that different feature selection methods select different features for the classification and the performance varies quite differently between the methods. The results also suggested that depending on the availability of the given number of attributes, different feature selections can be applied in consideration with the desirable performance. Our aim is to find out the best feature selection method for HD prediction in terms of the high performance achieved and the number of selected attributes, for this reason ILFS is preferable compared with other methods.

Feature selection plays a critical role in many real-life applications, especially in Healthcare diagnosis, through which doctors, clinicians and clinical experts could explore the most significant symptoms which drastically impact on the potential of having disease. In this study, we have successfully applied feature selection method based on data mining technique to apply in the application of HD prediction. For the 58 attributes provided, we have reduced and selected a subset of selected features and achieved the best HD prediction performance. Our method can be applied in many real-life applications or in other disease diagnosis applications to analyze the data, identify the risk factors to assist doctors in generating more accurate prediction. Our future work includes applying our method on a large variety of healthcare datasets (e.g. Breast Cancer, Chronic Kidney) and providing a more comprehensive analysis on the classification and feature selection methods.

REFERENCES

[1] Y.E. Shao, C.D. Hou, and C.C. Chiu, "Hybrid intelligent modeling schemes for heart disease classification" *Applied Soft Computing*, vol. 14, no. 1, pp. 47–52, 2014.

[2] R.D. Canlas, "Data mining in healthcare: Current applications and issues," *School of Information Systems & Management, Carnegie Mellon University, Australia*, 2009.

[3] Helma, Christoph, Eva Gottmann, and Stefan Kramer, "Knowledge discovery and data mining in toxicology" *Statistical methods in medical research*, vol. 9, no. 4, pp. 329–358, 2000.

[4] I-N. Lee, S-C. Liao, and M. Embrechts, "Data mining techniques applied to medical information", *Medical informatics and the Internet in medicine*, vol. 25, no. 2, pp 81–102, 2000.

[5] N. Deepika, K. Chandrashekar, "Association rule for classification of heart attack patients", *International Journal of Advanced Engineering Science and Technologies*, vol. 11, no. 2, pp. 253–57, 2011.

[6] K. Srinivas, B. Kavitha Rani, and Dr. A. Govrdhan,"Application of data mining techniques in healthcare and prediction of heart attacks", *International Journal on Computer Science and Engineering*, vol. 2, no. 2, pp. 250–255, 2011.

[7] A. Sudha, P. Gayathiri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods", *International Journal of Computer Applications*, vol. 43, no. 14, pp. 0975–8887, 2012.

[8] M. A. Jabbar, Priti Chandra, and B. L. Deekshatulu, "Cluster based association rule mining for heart attack prediction", *Journal of Theoretical and Applied Information Technology*, vol. 32, no. 2, pp. 196–201, 2011.

[9] Shouman, Mai, Tim Turner, and Rob Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients", *Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, 2012.

[10] Dangare, Chaitrali S., and Sulabha S. Apte, "Improved study of heart disease prediction system using data mining classification techniques", *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44–48, 2012.

[11] K. Usha Rani, "Analysis of heart diseases dataset using neural network approach", *International Journal of Data Mining and Knowledge Management Processive*, vol. 1, no. 5, pp. 1–8, 2011.

[12] Olatubosun Olabode and Bola Titilayo Olabode, "Cerebrovascular accident attack classification using multilayer feed forward artificial neural network with back propagation error", *Journal of Computer Science*, vol. 8, no. 1, pp.18–25, 2012.

[13] M. Anbarasi, E. Anupriya, and N.CH.S.N. Iyenga, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5370–5376, 2010.

[14] S. B. Patel, P. K. Yadav, D. D. Shukla, "Predict the diagnosis of heart disease patients using classification mining techniques", *IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS)*, vol. 4, no. 2, pp. 61–64, 2013.

[15] N. Singh, P. Ferozepur, S. Jindal, "Heart disease prediction using classification and feature selection techniques", *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, no. 2, 2018.

[16] H. Takci, "Improvement of heart attack prediction by the feature selection methods", *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 1, pp. 1–10, 2018.

[17] R. Suganya, S. Rajaram, A. S. Abdullah, V. Rajendran, "A novel feature selection method for predicting heart diseases with data mining techniques", *Asian Journal of Information Technology*, vol. 15, no. 8, 2016.

[18] M. Mirmozaffari, A. Alinezhad, A. Gilanpour, "Heart disease prediction with data mining clustering algorithms", *Int'l Journal of Computing, Communications & Instrumentation Engineering (IJCCIE)*, vol. 4, no. 1, 2017.

[19] K. Uma, M. Hanumathappa, "Heart disease prediction using classification techniques with feature selection method", *Adarsh Journal of Information Technology*, vol. 5, no. 2, pp. 22–29, 2016.

[20] Roffo, Giorgio, Melzi, Simone, Castellani, Umberto, Vinciarelli, Alessandro, "Infinite latent feature selection: a probabilistic latent graph-based ranking approach", *Computer Vision and Pattern Recognition*, 2017.

[21] http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names, The contents of the heart-disease directory.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines", *Mach. Learn.*, vol. 46, no. 1-3, pp. 389-422, 2002.

[23] D. Fehr, H. Veeraraghavan, A. Wibmer, T. Gondo, K. Matsumoto, HA. Vargas, E. Sala, H. Hricak, JO. Deasy, "Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images", *Proceedings of the National Academy of Sciences*, vol. 112, no. 46, E6265-73, 2015.

[24] T. Hofmann, "Probabilistic latent semantic analysis", *Proceedings of the Fifteenth conference on uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1999 (pp. 289-296).

[25] G. Roffo, S. Melzi, M. Cristani, "Infinite feature selection", *In Conf. IEEE International Conference on Computer Vision*, 2015 (pp. 4202-4210).

[26] https://www.mathworks.com/matlabcentral/fileexchange/56937-feature-selection-library, The MATLAB Feature Selection Library 2017.

[27] Jolliffe, Ian., "Principal component analysis", *International encyclopedia of statistical science*, Springer, Berlin, Heidelberg, 2011 (pp. 1094–1096).

[28] Ziasabounchi, Negar, Askerzade, Iman N., "A comparative study of heart disease prediction based on principal component analysis and clustering methods", *Turkish Journal of Mathematics and Computer Science (TJMCS)*, 16.17: 18, 2014.

[29] Hall, Mark Andrew, "Correlation-based feature selection for machine learning", *PhD thesis*, 1999.

[30] H. Zeng, Y. M. Cheung, "Feature selection and kernel learning for local learning-based clustering", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 33, no. 8, pp. 1532–1547, 2011.

Received on June 11, 2018
Revised on July 20, 2018