# An Efficient Hybrid Kernel Extreme Learning Machine Approach for Early Diagnosis of Parkinson's Disease

Hui-Ling Chen[1], Gang Wang[2], Chao Ma[3], Zhen-Nao Cai[4,1], Wen-Bin Liu[1], Su-Jing Wang[5*]

[1](College of Physics and Electronic Information, Wenzhou University, Wenzhou, Zhejiang, 325035, China)
[2](College of Computer Science and Technology, Jilin University, Changchun, 130012, China)
[3](School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen, 518172, China)
[4](School of Computer, Northwestern Polytechnical University, Xi'an, 710072, China)
[5](State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101,China)

**Abstract:**

In this paper, we explore the potential of extreme learning machine (ELM) and kernel ELM (KELM) for early diagnosis of Parkinson's disease (PD). In the proposed method, the key parameters including the number of hidden neuron and type of activation function in ELM, and the constant parameter $C$ and kernel parameter $\gamma$ in KELM are investigated in detail. With the obtained optimal parameters, ELM and KELM manage to train the optimal predictive models for PD diagnosis. In order to further improve the performance of ELM and KELM models, feature selection techniques are implemented prior to the construction of the classification models. The effectiveness of the proposed method has been rigorously evaluated against the PD data set in terms of classification accuracy, sensitivity, specificity and the area under the ROC (receiver operating characteristic) curve (AUC). Compared to the existing methods in previous studies, the proposed method has achieved very promising classification accuracy via 10-fold cross-validation (CV) analysis, with the highest accuracy of 96.47% and average accuracy of 95.97% over 10 runs of 10-fold CV.

***Keywords*:** Kernel extreme learning machine; Feature Selection; Medical diagnosis; Parkinson's disease

## 1. Introduction

Parkinson's disease (PD) is named after James Parkinson, who published the first paper describing this disease in 1817. Now PD has become the second most common degenerative disorders of the central nervous system after Alzheimer's disease [1]. It breaks out in large part of the world with fast rate, and the disease prevalence is expected to increase dramatically as the population ages [2], which might be particularly serious for developing countries such as China or India [3]. Till now, the cause of PD is still unknown, however, it is reported to be possible to alleviate symptoms

significantly at the onset of the illness in the early stage [4]. Patients with PD are usually characterized by five symptoms including tremor, rigidity, bradykinesia or slowness of movement, hand asymmetry and posture instability [5, 6]. Research has shown that approximately 90% of the patients with PD show vocal disorders [7]. It has also been proven that a vocal disorder may be one of the first symptoms to appear nearly 5 year before clinical diagnose [8]. The vocal impairment symptoms related with PD are known as dysphonia (inability to produce normal vocal sounds) and dysarthria (difficulty in pronouncing words) [9]. Little et al [10] has made use of the dysphonic indicators in their study to help discriminate PD patients from healthy ones. In their study, Support Vector Machine (SVM) with Gaussian kernel functions in combination with the feature selection approach was taken to predict PD, the simulation results have demonstrated that the proposed method can discriminate PD patients from healthy ones with approximately 90% classification accuracy using only four dysphonic features. A more recent study [11] has lifted the accuracy to 93% by increasing the number of features, and the result has been further boosted up to 99% through the use of a group of feature selection algorithms.

Motivated by the pioneer work in [10], many researchers made use of a comprehensive machine learning techniques to handle the PD diagnosis problem. In [12], Das presented a comparative study of using artificial neural networks (ANNs), DMneural, Regression and Decision Tree for effective diagnosis of PD, the experimental results have shown that ANNs yield the best results with the overall classification score of 92.9%. In [13], AStröm et al. proposed a parallel feed-forward neural network structure for diagnosis of PD, the highest classification accuracy of 91.20% was obtained. In [14], Sakar et al. used the mutual information based feature selection methods integrated with the SVM classifier for PD diagnosis, and the classification accuracy of 92.75% was achieved. In [15], Li et al. proposed a fuzzy-based non-linear transformation method in combination with the SVM classifier for prediction of PD, and the best classification accuracy of 93.47% was achieved. In [16], Shahbaba et al. introduced a new nonlinear model based on Dirichlet process mixtures for classification of PD, the results have been compared with that of multinomial logit models, decision trees, and SVM, the best classification accuracy of 87.7% was obtained by the proposed approach. In [17], Psorakis et al. introduced novel convergence measures, sample selection strategies and model improvements for multiclass mRVMs, and finally, the improved mRVMs achieved the classification accuracy rate of 89.47% when applied to prediction of PD. In [18], Guo et al. combined genetic programming and the expectation maximization algorithm (GP-EM) to detect PD, and the best classification accuracy of 93.1% was obtained. In [19], Luukka employed the feature selection method based on fuzzy entropy measures together with the similarity classifier to predict PD, and mean classification accuracy of 85.03%

with only two features was obtained. In [20], Ozcift et al. combined the correlation based feature selection (CFS) algorithm with the RF ensemble classifiers of 30 machine learning algorithms to identify PD, and the best classification accuracy of 87.13% was achieved by the proposed CFS-RF model. In [21], Spadoto et al. applied evolutionary-based techniques in combination with the Optimum-Path Forest (OPF) classifier to detect PD, and the best classification accuracy of 84.01% was achieved. In [22], Polat proposed to integrate the use of fuzzy c-means clustering-based feature weighting (FCMFW) with the k-NN classifier for the detection of PD, the classification accuracy of 97.93% was obtained. In [23], Chen et al. employed the Fuzzy k-nearest neighbour (FKNN) classifier in combination with the principle component analysis (PCA-FKNN) to diagnose PD, and the best classification accuracy of 96.07% was obtained by the proposed diagnosis system. In [24], Zuo et al. presented an effective and efficient diagnosis system based on particle swarm optimization enhanced FKNN for PD diagnosis, and the mean accuracy of 97.47% was reported. In [25], Hariharan et al. Developed a hybrid method by combining several feature pre-processing methods with classification techniques using least-square SVM, probabilistic neural network and general regression neural network, and the best classification accuracy of 100% was reported. In [26], Gök et al. developed a discriminative model by using rotation-forest ensemble k-nearest neighbour classifier algorithm, and the diagnosis accuracy of 98.46% was achieved.

From the above works, we can see that ANNs and SVM have gained much more popularity due to their mature theory background as well as the satisfactory classification performance. The main advantages of ANNs are their outstanding capability of capturing the nonlinearity relationship between the input and output existed in the data. However, it should be noted that the traditional gradient descent based training algorithm such as back propagation method may be easily trapped in the local minima as well as leaving many network parameters to be specified. Recently, Huang et al. proposed a new learning algorithm, extreme learning machine (ELM) [27], for a single hidden layer feed-forward neural networks (SLFNs). ELM chooses input weights and hidden biases randomly, and the output weights are analytically determined by using Moore–Penrose (MP) generalized inverse. However, one drawback of ELM is that the randomly assigned weights can produce a large variation in the classification accuracy in different trials. In order to solve this problem, more recently Huang et al. [28] proposed the kernelized version of ELM (KELM), which requiring no randomness in assigning connection weights between input and hidden layers. Compared with SVM, KELM can achieve comparative or better performance with much easier implementation and faster training speed in many classification or regression tasks [28-30].

Motivated by the excellent performance achieved by the ELM or KELM classifier on the

disease diagnosis problems such as thyroid disease diagnosis [31], erythemato-squamous diseases diagnosis [32] and paraquat-poisoned patients diagnosis [33], in this study, an attempt was made to explore the potential of ELM and KELM in constructing an automatic diagnostic system for diagnosis of PD. Previous study [10, 14, 15, 19, 23] on PD diagnosis have proven that using dimension reduction before conducting the classification task can improve the diagnosis accuracy. Here, an attempt is made to diagnose PD by using the ELM and KELM classifiers in combination with the feature selection methods. Four common feature selection techniques including maximum relevance minimum redundancy (mRMR), information gain (IG), Relief and t-test are employed for pre-processing before the classification models are constructed. The effectiveness of the proposed hybrid method is examined in terms of the classification accuracy, sensitivity, specificity and AUC on the PD data set taken from UCI machine learning repository. Promisingly, as can be seen that the developed method for this dataset in which a more reliable result is found (96.47% highest accuracy) over 10 runs of 10-fold cross validation (CV).

In summary, the main contributions of this paper can be summarized as follows: (1) The potential of ELM and KELM are explored in constructing an automatic diagnostic system for diagnosis of PD; (2) The detailed investigation on the impact of feature selection to the classification performance of PD diagnosis and interesting discovery are presented; (3) The most relevant measurement has been identified with the aid of the feature selection method.

The remainder of this paper is organized as follows. Section 2 offers brief background knowledge on ELM and KELM. In section 3 the detailed implementation of the proposed method is presented. Section 4 describes the experimental design. The experimental results and discussions of the proposed approach are presented in Section 5. Finally, Conclusions and recommendations for future work are summarized in Section 6.

## 2. Background Materials

## 2.1 ELM and KELM

This section gives a brief description of ELM. For more details, one can refer to [27, 34]. Given a training set $\aleph = \{(x_i, t_i) \mid x_i \in R^n, t_i \in R^m, i = 1, 2, \ldots, N\}$, where $x_i$ is the $n \times 1$ input feature vector and $t_i$ is a $m \times 1$ target vector. The standard SLFNs which has an activation function $g(x)$, and the number of hidden neurons $\tilde{N}$ can be mathematically modeled as follows:

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) = o_j,\ j = 1, 2, \ldots, N \qquad (1)$$

where $w_i$ is the weight vector between the *i*th neuron in the hidden layer and the input layer, $b_i$ means the bias of the *i*th neuron in the hidden layer; $\beta_i$ is the weight vector between the *i*th

hidden neuron and the output layer; and $o_j$ is the target vector of the *j*th input data. Here, $w_i \cdot x_j$ denotes the inner product of $w_i$ and $x_j$.

If SLFNs can approximate these N samples with zero error, we will have $\sum_{j=1}^{N} \| o_j - t_j \| = 0$, i.e., there exist $\beta_i$, $w_i$, $b_i$ such that $\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_i) = t_j,\ j = 1,2,\ldots,N$. The above Equation can be reformulated compactly as:

$$\mathrm{H}\beta = \mathrm{T} \tag{2}$$

where $$\mathrm{H}(w_1,\cdots,w_{\tilde{N}},b_1,\cdots,b_{\tilde{N}},x_1,\cdots,x_N) = \begin{pmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{pmatrix}_{N \times \tilde{N}} \tag{3}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m} \quad \text{and} \quad \mathrm{T} = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \tag{4}$$

As named by Huang et al.[35], H is called the hidden layer output matrix of the neural network, with the *i*th column of H being the *i*th hidden neuron output with respect to inputs $x_1, x_2, \ldots, x_N$. Huang et al. [36, 37] has shown that the input weights and the hidden layer biases of SLFNs need not be adjusted at all and can be arbitrarily given. Based on this assumption, the output weights can be analytically determined by finding the least square solution $\hat{\beta}$ of the linear system $\mathrm{H}\beta = \mathrm{T}$:

$$\| \mathrm{H}(w_1,\cdots,w_{\tilde{N}},b_1,\cdots,b_{\tilde{N}})\hat{\beta} - \mathrm{T} \| = \min_{\beta} \| \mathrm{H}(w_1,\cdots,w_{\tilde{N}},b_1,\cdots,b_{\tilde{N}})\beta - \mathrm{T} \| \tag{5}$$

Eq. (5) can be easily accomplished using a linear method, such as the Moor-Penrose (MP) generalized inverse of H, as is shown in Eq.(6)

$$\mathrm{H}\beta = \mathrm{T} \ \Rightarrow \ \hat{\beta} = \mathrm{H}^{\dagger}\mathrm{T} \tag{6}$$

where $\mathrm{H}^{\dagger}$ is the MP generalized inverse of the matrix H. The use of the MP generalized inverse method has led to the minimum norm least-squares (LS) solution, it is unique and has the smallest norm among all the LS solutions. As analyzed by Huang et al. [34], by using such MP inverse method, ELM tends to obtain a good generalization performance with a dramatically increased learning speed.

In summary, the learning steps of the ELM algorithm can be summarized as the following three steps:

Given a training set $\aleph = \{(x_i, t_i) \mid x_i \in R^n, t_i \in R^m, i = 1,2,\ldots,N\}$, an activation function $g(x)$, and the number of hidden neurons $\tilde{N}$,

(1) Randomly assign the input weights $w_i$ and bias $b_i$, $i = 1, 2, \ldots, \tilde{N}$.

(2) Calculate the hidden layer output matrix $\mathrm{H}$.

(3) Calculate the output weight $\beta = \mathrm{H}^{\dagger}\mathrm{T}$, $\mathrm{T} = [t_1, t_2, \ldots, t_n]^{\mathrm{T}}$.

It should be noted that when the feature mapping is unknown to users [28, 30], a kernel matrix for the ELM can be adopted according to the following equation:

$$\Omega_{ELM} = \mathrm{HH}^T : \Omega_{ELMi,j} = h(x_i) \cdot h(x_j) = K(x_i, x_j) \tag{7}$$

where $h(x)$ plays the role of mapping the data from the input space to the hidden-layer feature space $H$. The orthogonal projection method is adopted to calculate the Moore-Penrose generalized inverse of matrix, namely, $\mathrm{H}^{\dagger} = \mathrm{H}^T(\mathrm{HH}^T)^{-1}$, and a positive constant $C$ is added to the diagonal of $\mathrm{HH}^T$. Now we can write the output function of ELM as follows:

$$f(x) = h\beta = h(x)\mathrm{H}^T(\frac{I}{C} + \mathrm{HH}^T)^{-1}T = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_N) \end{bmatrix}^T \left(\frac{I}{C} + \Omega_{ELM}\right)^{-1} T \tag{8}$$

In this specific kernel implementation of ELM, namely KELM, we can specify the corresponding kernel for ELM model, the hidden layer feature mapping need not to be known to users. In this paper, the Gaussian radial basis function kernel $K(u, v) = \exp(-\gamma \|u - v\|^2)$ is applied. The two main parameters presented in KELM with Gaussian kernel are penalty parameter $C$ and kernel parameter $\gamma$, which play an important role in model construction. The parameter $C$ determines the trade-off between the fitting error minimization and the norm of input weights minimization, while the parameter $\gamma$ defines the non-linear mapping from the input space to some high-dimensional feature space.

## 2.2 Feature selection methods

Four common feature selection methods including maximum relevance minimum redundancy (mRMR), Information Gain (IG), Relief and t-test are adopted in this study.

### 2.2.1 Maximum relevance minimum redundancy (mRMR)

mRMR is a filter type feature selection method that seeks to choose features which are relevant to the target class (maximum relevance) and come up with the feature subset containing as non-redundant features as possible (minimum redundancy) [38]. It tries to determine the correlations between features and target class, features and features by using the mutual

information. The optimization criterion of mRMR is given as follows:

$$\max_{x_j \in X - S_{k-1}} \left[ I(x_j, c) - \frac{1}{k-1} \sum_{x_i \in S_{k-1}} I(x_j, x_i) \right] \tag{9}$$

where $X$ is the whole set of features, $c$ is the target class, $x_j$ is the $j$th feature, $S_{k-1}$ is the set of top $k$-1 features selected in the earlier iterations, $I$ is the mutual information, $I(x_j, c)$ and $I(x_i, x_j)$ denote mutual information between individual features $x_j$ with class $c$ and mutual information between features $x_i$ and $x_j$, respectively.

The mRMR feature selection framework attempts to select features based on a balance between maximizing the joint dependency of top ranking features on the target class and avoiding selecting redundant features.

## 2.2.2 Information Gain (IG)

IG is always used as a tool to measure the effectiveness of a feature in classifying instances. It is the change in information entropy from the prior uncertainty and expected posterior uncertainty using some feature [39], which is defined as:

$$IG = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y) \tag{10}$$

According to Eq.(10), IG is a symmetrical measure, where $H(Y)$ is the prior entropy and $H(Y|X)$ is the conditional entropy of the feature. It reflects additional information about $Y$ provides by $X$ that represents the amount by which the entropy of $Y$ decreases. The larger the value of IG, the more significant this feature is.

## 2.2.3 Relief

Relief is a measure of feature quality which is often used for feature subset selection. The idea of Relief is to reward the feature for having different values on a pair of similar examples from different classes, and punish it for having different values on examples from the same class [40].

For Relief algorithm, in each iteration, a sample $x$ is randomly selected and then two nearest neighbors of $x$ are found, one from the same classification (termed the nearest hit or NH) and the other from a dissimilar classification (termed the nearest miss or NM). So, Relief algorithm calculates the weight of the $i$th feature according to the following formulation:

$$w_i = w_i + | x^i - \mathrm{NM}^i(x) | - | x^i - \mathrm{NH}^i(x) | \tag{11}$$

where $w_i$ is the weight of the $i$th feature, $| x^i - \mathrm{NM}^i(x) |$ ($| x^i - \mathrm{NH}^i(x) |$) is the difference between the sample $x_i$ and its NM (NH) in the $i$th feature. That is, it may be a good feature if one sample has a large distance to its nearest neighbor sample from the dissimilar class, while it has a small distance to its nearest neighbor sample from the same class. Moreover, it is regarded as a real good feature when all samples support this rule. So, the $i$th feature is significant if $w_i$ is larger than a threshold, or it is not significant.

## 2.2.4 t-test

The t-test is often used to assess whether the means of two classes are statistically different from each other by calculating a ratio between the different of two class means and the variability of the two classes. It can be used commonly to determine the significance of each feature using the following equation [41]:

$$T_i = \frac{\mu_{c+}^i - \mu_{c-}^i}{\sqrt{\frac{(\sigma_{c+}^i)^2}{n_{c+}} + \frac{(\sigma_{c-}^i)^2}{n_{c-}}}} \quad (12)$$

where $\mu_{c+}^i$, $\mu_{c-}^i$, $\sigma_{c+}^i$, $\sigma_{c-}^i$ are the sample means and standard deviations in the *i*th feature of positive samples and negative samples, $n_{c+}$ and $n_{c-}$ are the size of positive samples and negative samples, respectively. The larger $T_i$ represents this feature is more significant.

## 3. Proposed hybrid method for PD diagnosis

The main objective of the proposed hybrid method is to provide an efficient and accurate diagnosis tool for PD diagnosis. The flowchart of the proposed ELM based and KELM based diagnosis method is shown in Figs. 1 and 2 respectively. In the proposed methods, feature selection is firstly applied to identify the informative features in PD dataset, after then several feature subsets with top ranked features are fed to the ELM and KELM model for performance evaluation. In Fig. 1, we can see that two main issues of the ELM based method are selection of hidden neurons and activation functions. While the main issue of the KELM based method is the choice of the parameter pair as shown in Fig. 2. The two hybrid methods are comprehensively evaluated on the PD dataset in terms AUC, ACC, sensitivity and specificity. The pseudo-code of the proposed method is given bellow.

---

**Pseudo-code for the proposed model**

/*Performance estimation by using *k*-fold CV where *k* = 10*/

**Begin**

**For** *j* = 1:*k*

Training set ← *k*-1 subsets;

Validation set ← remaining subset;

Rank features using mRMR, IG, Relief and t-test;

Train ELM and KELM classifiers on the reduced training data feature space using different size of feature subset;

Test the trained ELM and KELM models on the validation set;

**EndFor;**

Return the average classification accuracy rates of ELM and KELM over *j*th validation set;

**End.**

---

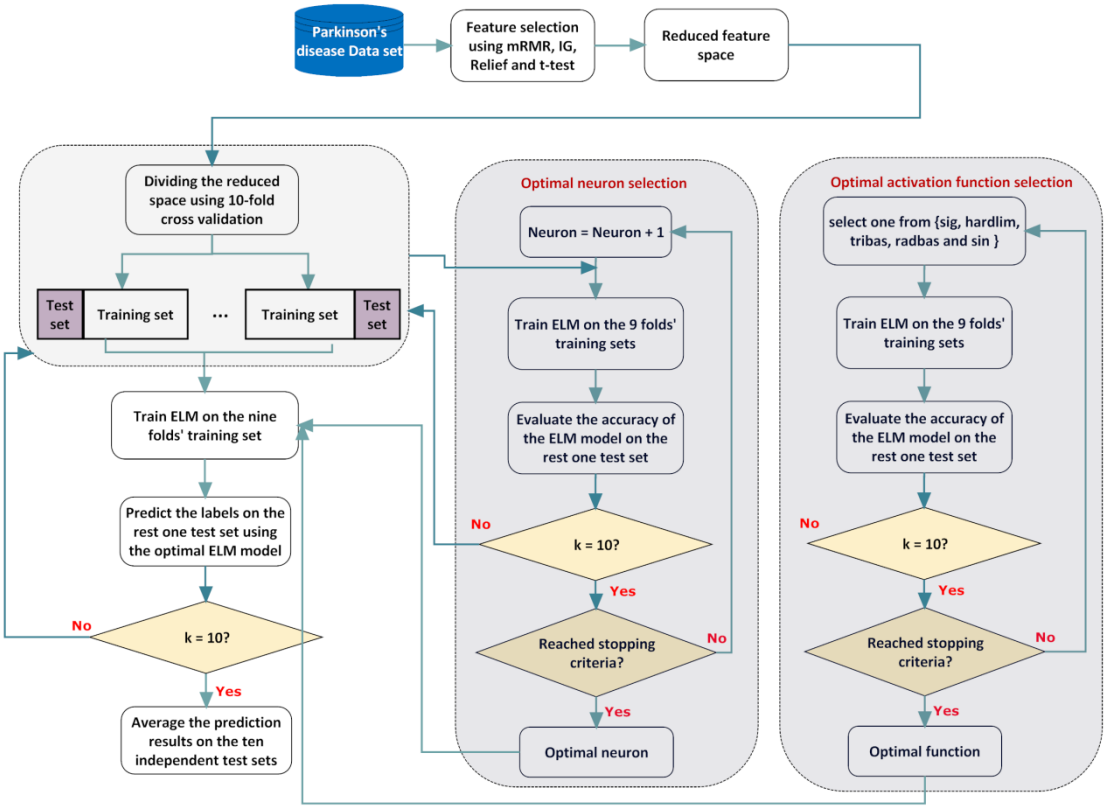



**Fig.1.** Overall procedure of the proposed ELM based diagnosis method.

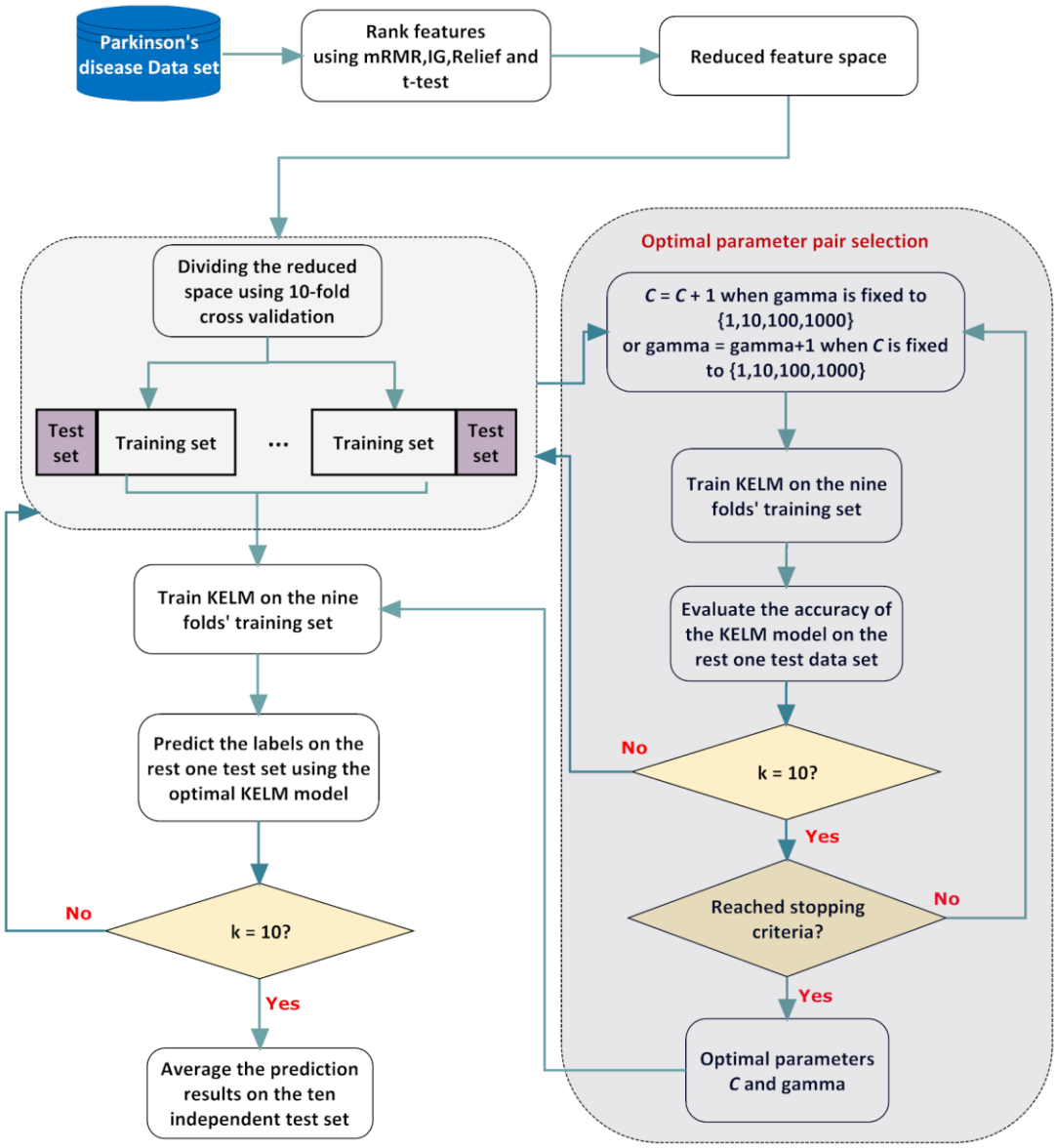



**Fig.2.** Overall procedure of the proposed KELM based diagnosis method.

# 4. Experiments design

## 4.1 Data Description

The experiment is conducted on the PD data set taken from UCI machine learning repository. (http://archive.ics.uci.edu/ml/datasets/Parkinsons, last accessed: May 2014). The purpose of this data set is to discriminate healthy people from those with PD, given the results of various medical tests carried out on a patient. This data set is composed of a range of biomedical voice measurements from 31 people, 23 with PD. The time since diagnoses ranged from 0 to 28 years, and the ages of the subjects ranged from 46 to 85 years, with a mean age of 65.8. Each subject provides an average of six phonations of the vowel (yielding 195 samples in total), each 36 seconds in length [42]. It should be noted that there is no missing values in the data set, and the whole features are real valued. The whole 22 features are presented in Table 1, along with its

description.

**Table 1** Description of the PD data set

| Label | Attribute | Description |
|---|---|---|
| F1 | MDVP:Fo(Hz) | Average vocal fundamental frequency |
| F2 | MDVP:Fhi(Hz) | Maximum vocal fundamental frequency |
| F3 | MDVP:Flo(Hz) | Minimum vocal fundamental frequency |
| F4 | MDVP:Jitter(%) | Several measures of variation in fundamental frequency |
| F5 | MDVP:Jitter(Abs) | |
| F6 | MDVP:RAP | |
| F7 | MDVP:PPQ | |
| F8 | Jitter:DDP | |
| F9 | MDVP:Shimmer | Several measures of variation in amplitude |
| F10 | MDVP:Shimmer(dB) | |
| F11 | Shimmer:APQ3 | |
| F12 | Shimmer:APQ5 | |
| F13 | MDVP:APQ | |
| F14 | Shimmer:DDA | |
| F15 | NHR | Two measures of ratio of noise to tonal components in the voice |
| F16 | HNR | |
| F17 | RPDE | Two nonlinear dynamical complexity measures |
| F18 | D2 | |
| F19 | DFA | Signal fractal scaling exponent |
| F20 | Spread1 | Three nonlinear measures of fundamental frequency variation |
| F21 | Spread2 | |
| F22 | PPE | |

**4.2 Experimental setup**

The whole experiment is conducted in the MATLAB platform, which runs on Windows 7 operating system with AMD Athlon 64 X2 Dual Core Processor 5000+ (2.6 GHz) and 4GB of RAM. mRMR program can be obtained from *http://penglab.janelia.org/proj/mRMR/index.htm*. The corresponding algorithms of IG and Relief from WEKA tool [43] are called by the main program which is implemented in MATLAB, and we implement the t-test from scratch. For ELM and KELM, the implementation by Huang available from *http://www3.ntu.edu.sg/home/egbhuang* is used.

It is difficult to compute the information entropies of the continuous features using a limited number of instances. Therefore, before using mRMR, IG, and Relief methods the continuous features are first discretized into multiple intervals using a supervised discretization method named MDL method [44]. After then, normalization is employed before classification, in order to avoid feature values in greater numerical ranges dominating those in smaller numerical ranges, as well as to avoid the numerical difficulties during the calculation. In this study, the data are scaled

into the interval of [0, 1] according to the Eq. (13), where $x$ is the original value, $x'$ is the scaled value, $max_a$ is the maximum value of feature a, and $min_a$ is the minimum value of feature $a$.

$$x' = \frac{x - min_a}{max_a - min_a} \tag{13}$$

In order to gain an unbiased estimate of the generalization accuracy, the $k$-fold CV was used to evaluate the classification accuracy [45]. This study set $k$ as 10, i.e., the data is divided into ten subsets. Each time, one of the 10 subsets is used as the test set and the remaining 9 subsets are put together to form a training set. Then the average error across all 10 trials is computed. The advantage of this method is that all of the test sets are independent and the reliability of the results could be improved. It should be pointed out that only one repetition of the 10-fold CV will not generate enough classification accuracies for comparison due to the arbitrariness partition of the data set. So the 10-fold CV will be repeated and averaged over 10 runs for accurate evaluation.

### 4.3 Performance Metric

Classification accuracy (ACC), sensitivity, specificity and AUC are commonly used as performance metrics for evaluation the performance of the binary classification task, especially for the task of disease diagnosis. In order to define these measures, the confusion matrix is introduced as shown in Table 2. Where TP is the number of true positives, which means that some cases with PD are correctly classified as ones with PD; FN, the number of false negatives, which means that some cases with PD are classified as healthy persons; TN, the number of true negatives, which means that some healthy persons are correctly classified as healthy persons; and FP, the number of false positives, which means that some healthy persons are classified as patients with PD.

**Table 2** The confusion matrix

| | Predicted patients with PD | Predicted healthy persons |
|---|---|---|
| Actual patients with PD | True Positive (TP) | False Negative (FN) |
| Actual healthy persons | False Positive (FP) | True Negative (TN) |

According to the confusion matrix, ACC, sensitivity and specificity are defined as follows:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \tag{14}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{15}$$

$$Specificity = \frac{TN}{FP + TN} \times 100\% \tag{16}$$

AUC represents the area under the receiver operating characteristic (ROC) curve, which plots

true positives rates versus the false positive rates. A classifier that provides a larger AUC is preferable over a classifier with a smaller AUC. A perfect classifier provides an AUC that equals 1. AUC is one of the best methods for comparing classifiers in two-class problems [46], in this study the method proposed in [47] was implemented to compute the AUC.

# 5 Experimental results and discussions

## 5.1 Experiment I: Classification in the Whole Original Feature Space

In this experiment, the performance of ELM and KELM for the PD diagnosis is examined. The performance of ELM is mainly influenced by the different types of activation functions and the number of hidden neurons. Here, these two key factors will be examined in detail. We firstly present results from our investigations on the influence of the different types of activation function and assign initial values for it. Five different common types of activation function including Sigmoid function (sig), Hard-limit function (hardlim), Triangular basis function (tribas), Radial basis function (radbas) and Sine function (sin) are investigated. The relationship between the classification accuracy of different ELM models and the different number of neurons is shown in Fig. 3. From Fig. 3 we can clearly see that the classification accuracy of ELM model is heavily influenced by the number of hidden neurons on the PD dataset. However, we can't obviously see which activation function performs best among them. Therefore, we further summarize the detailed results of ELM models on the PD dataset with five different activation functions by increasing the hidden neurons from 1 to 200 with the step of 1 in Table 3. All the results in Table 3 are shown in the form of average value (Mean), standard deviation (SD), maximum value (Max) and the minimum value (Min) over the all neurons. In Table 3, the most appropriate hidden neuron for each activation function is also recorded. It is found to be that ELM model achieves the best classification accuracy when the number of hidden neuron is set to be 57, 63, 84, 93 and 67 for sig, hardlim, tribas, radbas and sin activation function respectively.

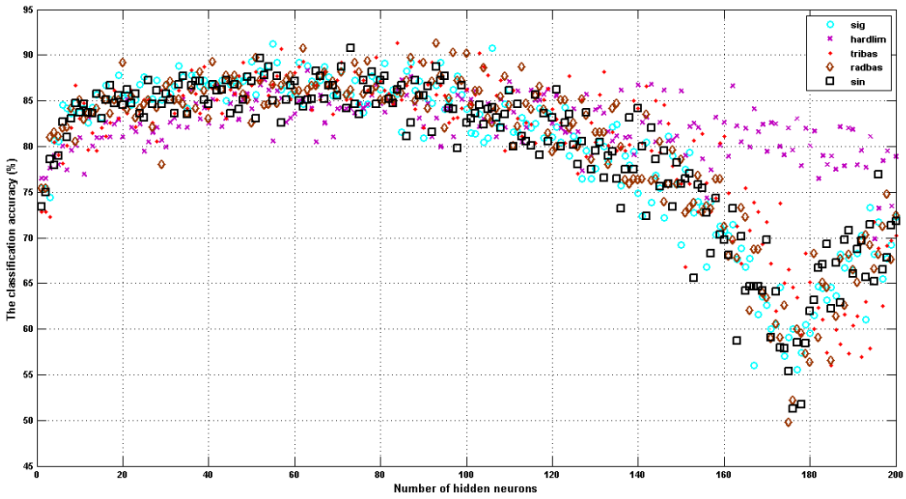



**Fig.3.** Trends of classification results of ELM with different activation functions by increasing the number of hidden neurons

**Table 3** Results of ELM with different activation functions on the PD dataset

| Type of activation function | Classification accuracy (%) | | | | Best hidden neuron |
|---|---|---|---|---|---|
| | Mean | Max | Min | SD | |
| sig | 79.95 | 91.29 | 51.29 | 8.74 | 57 |
| hardlim | 82.33 | 88.32 | 69.84 | 3.00 | 63 |
| tribas | 80.08 | 91.29 | 55.97 | 8.57 | 84 |
| radbas | 79.88 | 91.29 | 49.76 | 8.89 | 93 |
| sin | 78.79 | 90.32 | 53.21 | 9.09 | 67 |

The detailed results of 10 runs of 10-fold CV of ELM models with different activation functions by taking the acquired best hidden neuron are summarized in Fig. 4 and Table 4. From the table, we can see that ELM with the Sine function outperforms ELM with other activation functions with the average accuracy of 86.61%, the maximum accuracy of 89.79% and the SD of 1.67% over 10 runs of 10-fold CV. It is interesting to find that the standard deviation obtained by the ELM with Sine function is the smallest among the five activation functions. It indicates that ELM with Sine function is much more stable than other ELM models. Therefore, the Sine function is adopted in the subsequent experiment analysis.

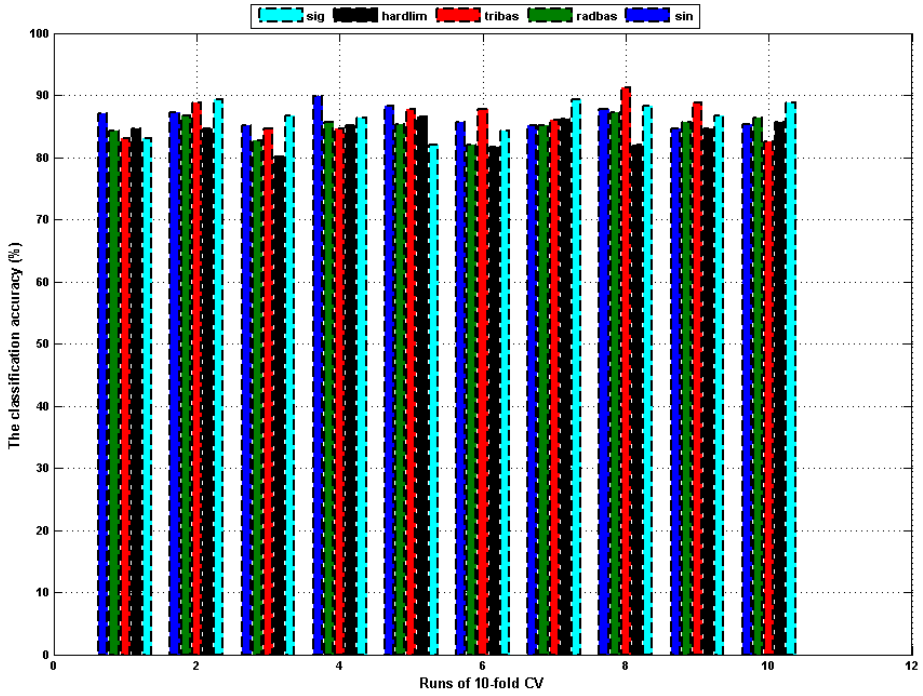



**Fig.4.** Classification results of 10 runs of 10-fold CV of ELM models with different activation functions

**Table 4** Results of 10 runs of 10-fold CV of different ELM models

| Type of activation function | Classification accuracy (%) | | | |
|---|---|---|---|---|
| | Mean | SD | Max | Min |
| sig | 86.46 | 2.59 | 89.29 | 82.08 |
| hardlim | 84.11 | 2.12 | 86.50 | 80.13 |
| tribas | 86.51 | 2.80 | 91.26 | 82.58 |
| radbas | 85.08 | 1.67 | 87.16 | 82.05 |
| sin | **86.61** | **1.67** | **89.79** | **84.68** |

Different from ELM, the performance of KELM is mainly influenced by the constant $C$ and kernel parameter $\gamma$ in Gaussian kernel function. Therefore, the impact of these two parameters on KELM model for PD diagnosis is also examined in detail in this experiment. In order to investigate the impacts of these parameters, we have conducted the experiments using different values of $C$ when the value of $\gamma$ is fixed to 1, 10, 100 and1000 respectively, different values of $\gamma$ when the value of $C$ is fixed to 1, 10, 100 and 1000 respectively. The relationship between classification accuracy and parameter $C$ with different values of $\gamma$, and parameter $\gamma$ with different values of $C$ are shown in Figs. 5 and 6 respectively. From Fig. 5 we can clearly see that parameter

$\gamma$ has a big impact to the performance of KELM classifier. Interestingly, the classification accuracy is getting higher when the value of $\gamma$ is set to be smaller. The best classification accuracy of 89.79%, 91.26%, 93.87% and 94.89% is achieved with the parameter pair of (1, 1), (10, 1), (100, 2) and (1000, 2) as shown in Fig. 5(a), Fig. 5(b), Fig. 5(c) and Fig. 5(d) when $C$ is equal to 1, 10,100 and 1000 respectively. Compared to the parameter $\gamma$, the parameter $C$ is not sensitive to the performance of KELM. From Fig. 6 we can see that the classification accuracy is fluctuating when changing the value of $C$. The best classification accuracy of 96.45%, 89.82%, 87.68% and 86.17% is achieved with the parameter pair of (62, 1), (84, 10), (94, 100) and (48, 1000) as shown in Fig. 6(a), Fig. 6(b), Fig. 6(c) and Fig. 6(d) when parameter $\gamma$ is equal to 1, 10,100 and 1000 respectively. Owing to the best classification accuracy is achieved when $C$ and $\gamma$ is set to be 62 and 1 respectively, the optimal parameter pair of (62, 1) is adopted for subsequent analysis.

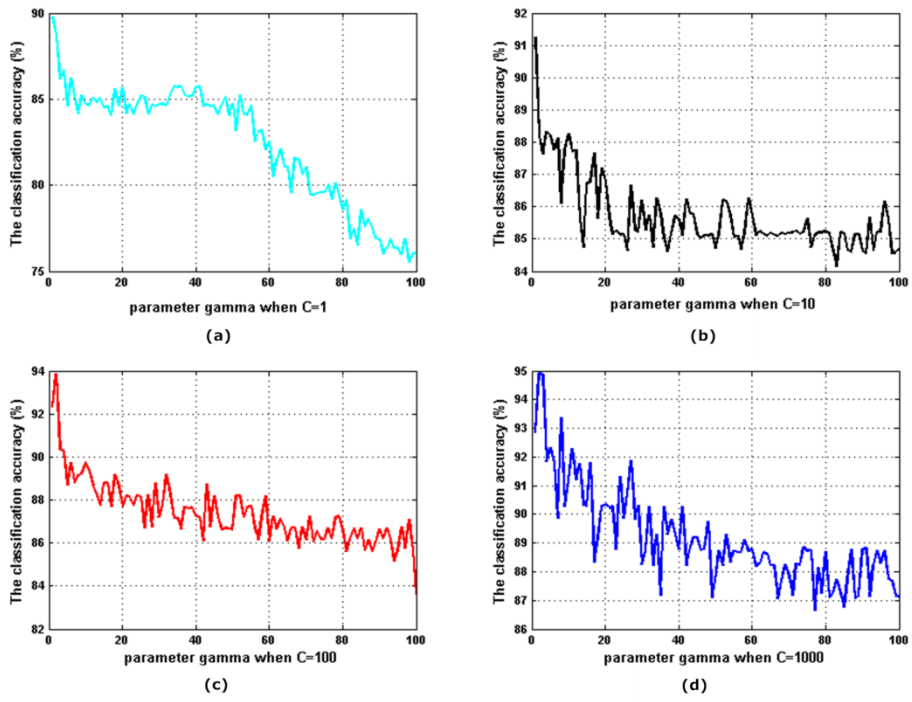



**Fig.5.** The relationship between classification accuracy and parameter $\gamma$ with different values of parameter $C$

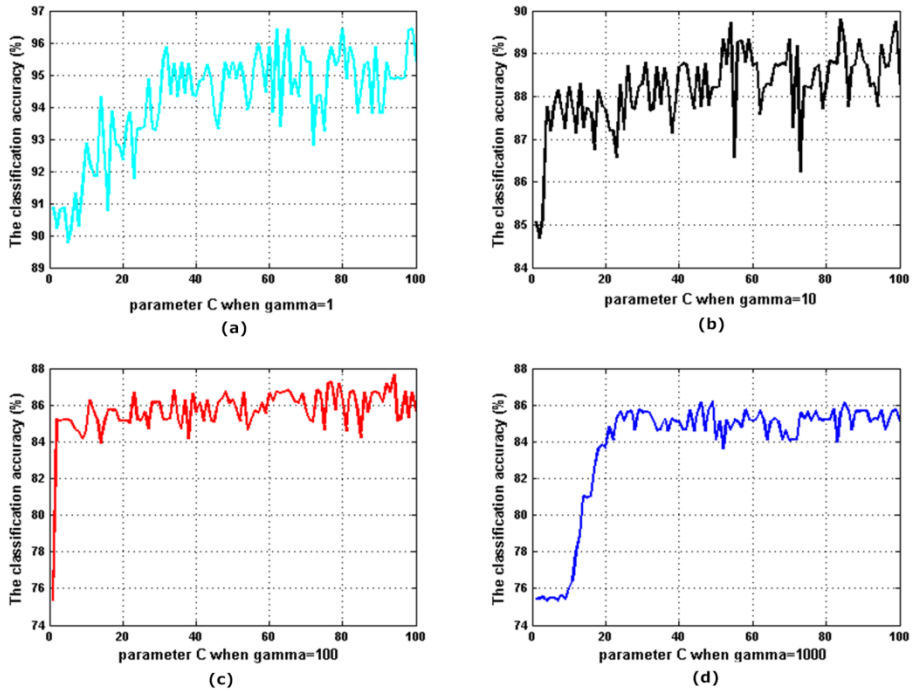



**Fig.6.** The relationship between classification accuracy and parameter *C* with different values of parameter $\gamma$.

The detailed results of 10 runs of 10-fold CV of KELM models with the optimal parameter pair are listed in Table 5. From the table, we can see that KELM model achieves high performance with average results of 92.85%, 94.63%, 96.93% and 88.78% in terms of AUC, ACC, sensitivity and specificity respectively. Compared with the best ELM model, KELM with the optimal parameter pair has achieved the average classification accuracy with an increase from 86.61% to 94.63%, the boosted 8% classification accuracy obtained by the KELM model may be owing to the fact that the constructed KELM model is able to effectively capture the nonlinear relationship existed in the PD dataset with the aid of the Gaussian kernel. In addition, the acquired standard deviation of KELM is also much smaller than that of ELM model. It also indicates the stability and robustness of the KELM model.

**Table 5** The detailed results obtained by KELM model with optimal parameter pair.

| Runs of 10-fold CV | AUC | ACC | Sensitivity | Specificity |
|---|---|---|---|---|
| #1 | 0.9456 | 0.9492 | 0.9695 | 0.9217 |
| #2 | 0.9181 | 0.9489 | 0.9803 | 0.8558 |
| #3 | 0.9312 | 0.9432 | 0.9752 | 0.8871 |
| #4 | 0.9435 | 0.9537 | 0.9729 | 0.9142 |
| #5 | 0.9164 | 0.9387 | 0.9661 | 0.8667 |
| #6 | 0.9299 | 0.9489 | 0.9665 | 0.8933 |
| #7 | 0.9457 | 0.9545 | 0.9647 | 0.9267 |
| #8 | 0.9206 | 0.9379 | 0.9597 | 0.8814 |
| #9 | 0.8982 | 0.9339 | 0.9707 | 0.8256 |
| #10 | 0.9360 | 0.9545 | 0.9671 | 0.9050 |
| Avg. | **0.9285** | **0.9463** | **0.9693** | **0.8878** |
| Dev. | 0.0153 | 0.0075 | 0.0058 | 0.0317 |

In order to evaluate the effectiveness of the proposed KELM approach, the SVM was also implemented for comparison. Here we considered both the linear kernel (SVM_Linear) and nonlinear RBF kernel (SVM_RBF) for SVM classification. For SVM_Linear, the penalty parameter $C$ was chosen from the set of {0.01, 0.1, 1, 10, 100, 1000}. According to the preliminary analysis, the best classification performance of SVM_Linear was achieved when the value of $C$ was set to be 1000. For SVM_RBF, a grid-search technique [48] was used to obtain the optimal parameter values of RBF kernel function. The range of the related parameters $C$ and $\gamma$ were varied between $C = \{2^{-5},2^{-3},\ldots,2^{15}\}$ and $\gamma = \{2^{-15},2^{-13},\ldots,2^{1}\}$. The optimal parameter pair $(C,\gamma)$ was employed to construct the predictive model. The detailed classification performance of SVM_Linear and SVM_RBF with the maximum value (Max), minimum value (Min), average value (Mean) and the standard value (SD) over 10 runs of 10-fold CV are recorded in Table 6. As can be seen from the table, the average classification accuracy obtained by the SVM_Linear and SVM_RBF is 87.30% and 93.34%, respectively. The SVM model with RBF kernel has achieved much better performance than the model with linear kernel. It indicates that the nonlinear relationship between the features of the PD data is well captured by the SVM_RBF model. However, when the SVM_RBF is compared with the KELM approach, we can find that the KELM performs even better than SVM_RBF. From the Tables 5 and 6, we can see that the average ACC and AUC of KELM are higher than that of SVM_RBF by 1.29% and 2.8%, respectively. In addition, we can see that the standard deviation of KELM is also smaller than that of the SVM models, which indicates the consistency and stability of the KELM model.

**Table 6** The detailed results obtained by SVM model.

| Performance Metric | Classification performance | | | |
|---|---|---|---|---|
| | Mean | SD | Max | Min |
| ACC(SVM_RBF) | 0.9334 | 0.0142 | 0.9589 | 0.9137 |
| AUC(SVM_RBF) | 0.9005 | 0.0292 | 0.9416 | 0.8565 |
| Sensitivity(SVM_RBF) | 0.8342 | 0.0496 | 0.9033 | 0.7577 |
| Specificity(SVM_RBF) | 0.9669 | 0.0112 | 0.9798 | 0.9475 |
| ACC(SVM_Linear) | 0.8730 | 0.0106 | 0.8926 | 0.8571 |
| AUC(SVM_Linear) | 0.8188 | 0.0207 | 0.8481 | 0.7935 |
| Sensitivity(SVM_Linear) | 0.7065 | 0.0342 | 0.7493 | 0.6549 |
| Specificity(SVM_Linear) | 0.9311 | 0.0117 | 0.9469 | 0.9132 |

### 5.2 Experiment II: Classification with Feature Selection

To investigate whether feature selection can further improve the performance of ELM and KELM for diagnosis of PD, we further conduct the experiments in the reduced feature space. mRMR, IG, Relief and t-test are implemented to rank the features and the trends of classification accuracy of ELM and KELM model over the incremental feature subset are shown in Fig. 7. For convenience, the hidden neuron of 67 is taken for ELM model with Sine function, and the parameter pair of (1, 62) is adopted for KELM. From Fig. 7 we can see that feature selection can further improve the classification accuracy of the ELM and KELM, except the IG approach. Both ELM and KELM combined with IG achieve the best performance with the feature subset be full with the whole 22 features. It can be also found that the two models coupled with mRMR filter achieve the best classification accuracy with the smallest features among the four feature selectors. Therefore, mRMR has emerged as the promising technique compared to other three feature selection methods for extracting most informative features. In addition, we can find that KELM still performs much better than ELM with the aid of feature selection.

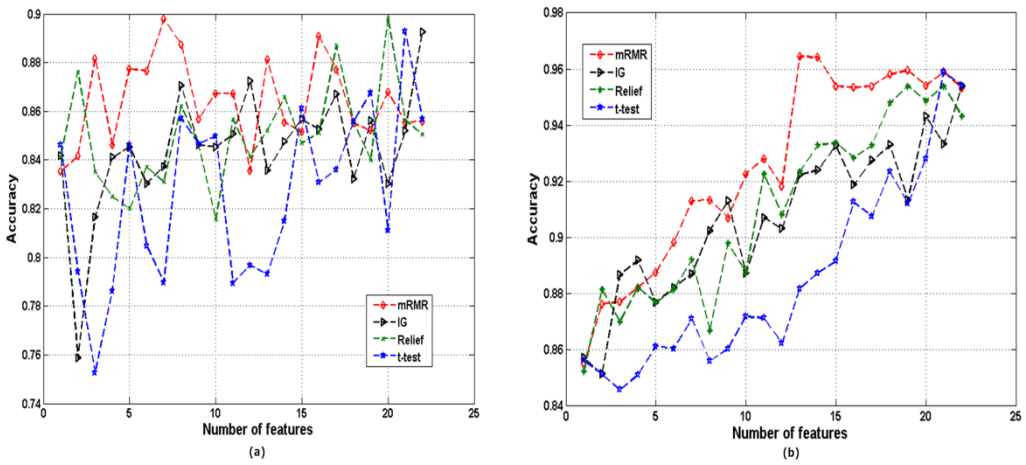


**Fig.7.** Trends of classification accuracy of ELM and KELM for different feature subset obtained by different feature selection methods: (a) ELM model (b) KELM model.

Since both ELM and KELM are sensitive to the variation of the parameter values on different feature subset, further detailed evaluation should be conducted. For simplicity, here we performed the detailed evaluation for KELM model with the mRMR filter owing to its excellent discriminative ability. We first utilize mRMR to rank the features and then selected top 1, 5, 10, 15, and 20 features as shown in Table 7. Since KELM model is sensitive to the variation of the parameter $C$ and $\gamma$, we performed the experiment to look for the best parameter pair in each feature subset. Fig. 8 shows the trends of classification accuracy for different feature subset by changing the value of parameter $C$ in the range of [1, 100] at the step size of 1 when the value of parameter $\gamma$ is fixed to 1. From Fig. 8 we can clearly see that KELM gets different classification results on different feature subsets, and the trends of classification accuracy seems to be increasing when the size of feature subset is growing. From this figure, we have also got the best parameter pair for KELM on each feature subset. The best parameter pair for feature subset size 1, 5, 10, 15 and 20 is (9, 1), (92, 1), (85, 1), (86, 1) and (84, 1), respectively. These parameter pairs are adopted for the subsequent experimental analysis. Table 8 lists the detailed results of KELM construed on different feature subsets in terms of AUC, ACC, sensitivity and specificity. From Table 8 we can observe the following facts:

1) The performance of KELM models built with feature subset size of 15 and 20 is better than the one built with all features. The best performance of KELM is obtained on the feature subset with size of 20, with the average AUC of 94.37%, ACC of 95.97%, sensitivity of 97.61% and specificity of 91.12%.
2) Among six feature subset sizes, the results show that the size of 15 is enough to build classification model, the KELM model with feature subset size of 15 achieves the average AUC of 94.19%, ACC of 95.49%, sensitivity of 97.27% and specificity of 91.11%, which is better than those obtained by using all features.
3) The sensitivity of all models is close to each other, KELM can achieve the sensitivity of 93.20% using only one feature. It indicates the first feature PPE, a nonlinear measure of fundamental frequency variation, selected by mRMR filter is one of the most informative feature, this result is consistent with the earlier finding obtained in [10] .

**Table 7** The feature subset obtained by mRMR filter.

| Size | Feature subset |
|---|---|
| 1 | F22 |
| 5 | F22 F18 F1 F13 F20 |
| 10 | F22 F18 F1 F13 F20 F15 F3 F2 F6 F21 |
| 15 | F22 F18 F1 F13 F20 F15 F3 F2 F6 F21 F19 F12 F17 F10 F5 |
| 20 | F22 F18 F1 F13 F20 F15 F3 F2 F6 F21 F19 F12 F17 F10 F5 F8 F9 F7 F11 F4 |

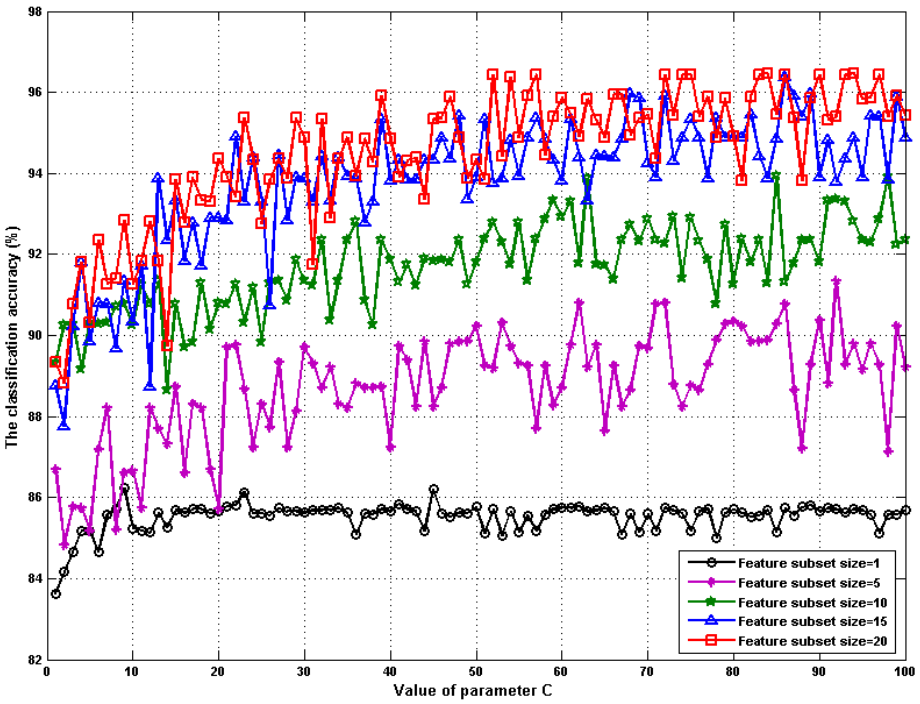



**Fig.8.** Trends of classification accuracy for different feature subset by changing the value of parameter *C* when parameter $\gamma$=1.

**Table 8** Performance of KELM using different feature subsets.

| Feature subset | AUC | ACC | Sensitivity | Specificity | Optimal parameter pair |
|---|---|---|---|---|---|
| 1 | 0.7689 [0.0188] | 0.8562 [0.0017] | 0.9320 [0.0021] | 0.6059 [0.0368] | (9,1) |
| 5 | 0.8436 [0.0151] | 0.8988 [0.0087] | 0.9571 [0.0087] | 0.7301 [0.0235] | (92,1) |
| 10 | 0.8845 [0.0226] | 0.9283 [0.0114] | 0.9679 [0.0077] | 0.8012 [0.0426] | (85,1) |
| 15 | 0.9419 [0.0141] | 0.9549 [0.0061] | 0.9727 [0.0067] | 0.9111 [0.0274] | (86,1) |
| 20 | 0.9437 [0.0076] | 0.9597 [0.0050] | 0.9761 [0.0076] | 0.9112 [0.0140] | (84,1) |
| All features | 0.9285 [0.0153] | 0.9463 [0.0075] | 0.9693 [0.0058] | 0.8878 [0.0317] | (62,1) |

Value inside the square brackets shows standard deviation of 10 runs of 10-fold CV.

From the above analysis, we can find that with the aid of feature selection using mRMR, KELM has improved its performance for PD diagnosis in terms of AUC, ACC, sensitivity and specificity. In addition, it is interesting to find that the standard deviation of KELM is becoming smaller than before in most cases, which indicates that KELM has become more robust and reliable through feature selection. Table 9 also presents the optimal confusion matrices obtained by KELM models over the 10 runs of 10-fold CV with different feature subsets. As can be seen from Table 9, KELM with feature subset size of 15 and 20 can correctly classify 144 PD cases out of 147 total PD cases, while misclassify 3 patients with PD as healthy persons and 4 cases of healthy persons as patients with PD. While KELM with the whole features correctly classifies 143 PD cases out of 147 total PD cases, misclassifies 5 patients with PD as healthy persons and 4 healthy persons as patients with PD.

**Table 9** Optimal Confusion matrix of KELM using different feature subsets.

| Feature subset | | Predicted patients with PD | Predicted healthy persons |
|---|---|---|---|
| 1 | Actual patients with PD | **137** | 10 |
| | Actual healthy persons | 18 | **30** |
| 5 | Actual patients with PD | **142** | 5 |
| | Actual healthy persons | 13 | **35** |
| 10 | Actual patients with PD | **144** | 3 |
| | Actual healthy persons | 8 | **40** |
| 15 | Actual patients with PD | **144** | 3 |
| | Actual healthy persons | 4 | **44** |
| 20 | Actual patients with PD | **144** | 3 |
| | Actual healthy persons | 4 | **44** |
| All features | Actual patients with PD | **143** | 5 |
| | Actual healthy persons | 4 | **43** |

To show the trends of the classification performance of KELM and ELM over the different feature space, KELM and ELM with different parameter values are implemented. For convenience, the hidden neurons of ELM are set to be 10, 50 and 100, and they are named ELM1, ELM2 and ELM3 respectively. The parameter pair for KELM are set to be (1, 10), (1, 50) and (1, 100), and they are named KELM1, KELM2 and ELM3 respectively. Fig. 9 shows the comprehensive results obtained by the KELM and ELM classifiers in terms of ACC, AUC, sensitivity and specificity in one run of 10-fold CV on the reduced feature space where the ranked features obtained by mRMR range from 1 to 22 with the step size of 1. It can be observed that KELM achieves the better results than ELM in terms of ACC, AUC, sensitivity and specificity on the reduced space in most cases. However, the sensitivity obtained by ELM1 is very close to that of KELM models. It means that ELM with the hidden neuron of 10 can achieve the same ability to discriminate the patients with PD as that of KELM.

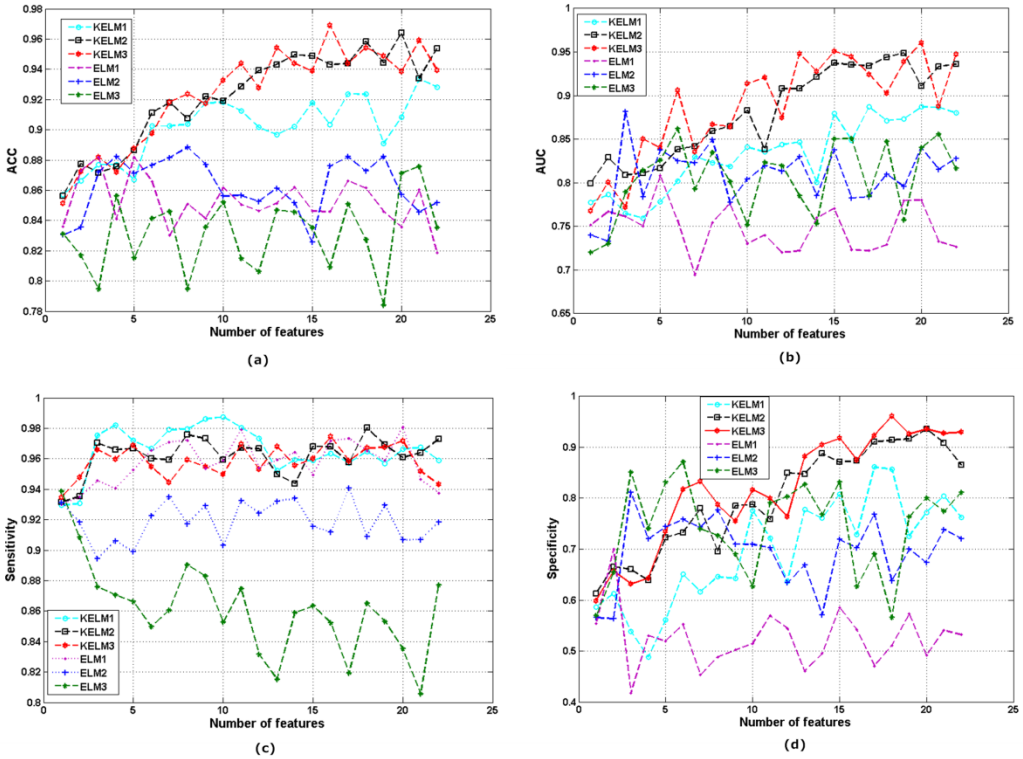



**Fig.9.** Trends of the classification performance over different reduced feature space.

For comparison purpose, the classification accuracies obtained by the previous methods on the same the PD dataset are listed in Table 10. It can be seen that our developed approach has achieved promising results with the highest accuracy of 96.47% and mean accuracy of 95.97%. The promising performance of the proposed hybrid method might be very helpful in assisting the physicians to make the accurate diagnosis on the patients and will show great potential in the area of clinical PD diagnosis.

**Table 10** Classification accuracies obtained with our method and other methods

| Study | Method | Accuracy (%) |
|---|---|---|
| Little et al. (2009) | Pre-selection filter + Exhaustive search + SVM | 91.4(bootstrap with 50 replicates) |
| Shahbaba et al. (2009) | Dirichlet process mixtures | 87.7(5-fold CV) |
| Das (2010) | ANN | 92. (hold-out) |
| Sakar et al. (2010) | Mutual information based feature selection + SVM | 92.75(bootstrap with 50 replicates) |
| Psorakis et al. (2010) | Improved mRVMs | 89.47(10-fold CV) |
| Guo et al. (2010) | GP-EM | 93.1(10-fold CV) |
| Ozcift et al. (2011) | CFS-RF | 87.1(10-fold CV) |
| Li et al. (2011) | Fuzzy-based non-linear transformation + SVM | 93.47(hold-out) |
| Luukka (2011) | Fuzzy entropy measures + Similarity classifier | 85.03(hold-out) |
| Spadoto et al. (2011) | Particle swarm optimization + OPF | 73.53(hold-out) |
| | Harmony search + OPF | 84.01(hold-out) |
| | Gravitational search algorithm + OPF | 84.01(hold-out) |
| AStröm et al. (2011) | Parallel NN | 91.20(hold-out) |
| Chen et al. (2013) | PCA-FKNN | 96.07 (average 10-fold CV) |
| This Study | mRMR-KELM | 95.97(average 10-fold CV) |
| | | 96.47(10-fold CV) |

## 6. Conclusions and future works

In this work, we have developed an efficient hybrid method, mRMR-KELM, for addressing PD diagnosis problem. The core component of the proposed method is the KELM classifier, whose key parameters are explored in detail. With the aid of the feature selection techniques, especially the mRMR filter, the performance of KELM classifier is ameliorated with much smaller features. The promising performance obtained on the PD dataset has proven that the proposed hybrid method can distinguish well enough between patients with PD and healthy persons. It is observed that mRMR-KELM achieves the highest classification accuracy of 96.47% via 10-fold CV analysis. Based on the empirical analysis, it can be safely concluded that, the developed diagnosis method can assist the physicians to make accurate diagnostic decision. The future investigation will pay much attention to evaluating the proposed method in other medical diagnosis problems.

## Acknowledgements

This research is supported by the National Natural Science Foundation of China (61303113, 61379095, 61272018, 61402337), the Zhejiang Provincial Natural Science Foundation of China under Grant Nos. R1110261, LY14F020035, LQ13G010007 and LQ13F020011, the China Postdoctoral Science Foundation funded project (2012M520428, 2014T70133), the Open Projects Program of National Laboratory of Pattern Recognition (201306295), and the Beijing Natural Science Foundation (4152055).

## References

1. de Lau, L.M.L. and M.M.B. Breteler, *Epidemiology of Parkinson's disease.* The Lancet Neurology, 2006. **5**(6): p. 525-535.
2. Van Den Eeden, S.K., et al., *Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity.* American Journal of Epidemiology, 2003. **157**(11): p. 1015-1022.
3. Dorsey, E., et al., *Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030.* Neurology, 2007. **68**(5): p. 384-386.
4. Singh, N., V. Pillay, and Y.E. Choonara, *Advances in the treatment of Parkinson's disease.* Progress in neurobiology, 2007. **81**(1): p. 29-44.
5. Jankovic, J., *Parkinson's disease: clinical features and diagnosis.* Journal of Neurology, Neurosurgery & Psychiatry, 2008. **79**(4): p. 368-376.
6. Massano, J. and K.P. Bhatia, *Clinical approach to Parkinson's disease: features, diagnosis, and principles of management.* Cold Spring Harbor Perspectives in Medicine, 2012. **2**(6).
7. Ho, A.K., et al., *Speech impairment in a large sample of patients with Parkinson's disease.* Behavioural neurology, 1998. **11**: p. 131-138.
8. Harel, B., M. Cannizzaro, and P.J. Snyder, *Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study.* Brain and Cognition, 2004. **56**(1): p. 24-29.
9. Baken, R.J. and R.F. Orlikoff, *Clinical measurement of speech and voice (2nd ed.)*. 2000, Singular Publishing Group, San Diego, CA.
10. Little, M.A., et al., *Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease.* Biomedical Engineering, IEEE Transactions on, 2009. **56**(4): p. 1015-1022.
11. Tsanas, A., et al., *Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease.* Biomedical Engineering, IEEE Transactions on, 2012. **59**(5): p. 1264-1271.
12. Das, R., *A comparison of multiple classification methods for diagnosis of Parkinson disease.* Expert Systems with Applications, 2010. **37**(2): p. 1568-1572.
13. AStröm, F. and R. Koker, *A parallel neural network approach to prediction of Parkinson's Disease.* Expert Systems with Applications, 2011. **38**(10): p. 12470-12474.
14. Sakar, C.O. and O. Kursun, *Telediagnosis of Parkinson's Disease Using Measurements of Dysphonia.* Journal of Medical Systems, 2010. **34**(4): p. 1-9.
15. Li, D.C., C.W. Liu, and S.C. Hu, *A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets.* Artificial intelligence in medicine, 2011. **52**(1): p. 45-52.
16. Shahbaba, B. and R. Neal, *Nonlinear models using Dirichlet process mixtures.* The Journal of Machine Learning Research, 2009. **10**: p. 1829-1850.
17. Psorakis, I., T. Damoulas, and M.A. Girolami, *Multiclass Relevance Vector Machines: Sparsity and Accuracy.* Neural Networks, IEEE Transactions on, 2010. **21**(10): p. 1588-1598.
18. Guo, P.F., P. Bhattacharya, and N. Kharma, *Advances in Detecting Parkinson's Disease.* Medical Biometrics, 2010: p. 306-314.
19. Luukka, P., *Feature selection using fuzzy entropy measures with similarity classifier.* Expert Systems with Applications, 2011. **38**(4): p. 4600-4607.
20. Ozcift, A. and A. Gulten, *Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms.* Comput Methods Programs Biomed, 2011. **104**(3): p. 443-451.

21. Spadoto, A.A., et al. *Improving Parkinson's disease identification through evolutionary-based feature selection*. in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*. 2011.
22. Polat, K., *Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering.* International Journal of Systems Science, 2012. **43**(4): p. 597-609.
23. Chen, H.-L., et al., *An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach.* Expert Systems with Applications, 2013. **40**(1): p. 263-271.
24. Zuo, W.-L., et al., *Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach.* Biomedical Signal Processing and Control, 2013. **8**(4): p. 364-373.
25. Hariharan, M., K. Polat, and R. Sindhu, *A new hybrid intelligent system for accurate detection of Parkinson's disease.* Computer Methods and Programs in Biomedicine, 2014. **113**(3): p. 904-913.
26. Gok, M., *An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease.* International Journal of Systems Science, 2015. **46**(6): p. 1108-1112.
27. Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew, *Extreme learning machine: Theory and applications.* Neurocomputing, 2006. **70**(1-3): p. 489-501.
28. Huang, G.B., et al., *Extreme Learning Machine for Regression and Multiclass Classification.* Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2012. **42**(2): p. 513-529.
29. Pal, M., A.E. Maxwell, and T.A. Warner, *Kernel-based extreme learning machine for remote-sensing image classification.* Remote Sensing Letters, 2013. **4**(9): p. 853-862.
30. Cheng, C., W.P. Tay, and G.-B. Huang. *Extreme learning machines for intrusion detection*. in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. 2012. IEEE.
31. Li, L.N., et al., *A Computer Aided Diagnosis System for Thyroid Disease Using Extreme Learning Machine.* Journal of medical systems, 2012. **36**(5): p. 3327-3337.
32. Liu, T., et al., *A fast approach for detection of erythemato-squamous diseases based on extreme learning machine with maximum relevance minimum redundancy feature selection.* International Journal of Systems Science, 2015. **46**(5): p. 919-931.
33. Hu, L., et al., *An efficient machine learning approach for diagnosis of paraquat-poisoned patients.* Computers in Biology and Medicine, 2015. **59**(0): p. 116-124.
34. Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew, *Extreme learning machine: a new learning scheme of feedforward neural networks*, in *IEEE International Joint Conference on Neural Networks.* 2004. p. 985-990.
35. Huang, G.B. and H.A. Babri, *Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions.* Neural Networks, IEEE Transactions on, 1998. **9**(1): p. 224-229.
36. Huang, G.B., *Learning capability and storage capacity of two-hidden-layer feedforward networks.* Neural Networks, IEEE Transactions on, 2003. **14**(2): p. 274-281.
37. Huang, G.B., L. Chen, and C.K. Siew, *Universal approximation using incremental constructive feedforward networks with random hidden nodes.* Neural Networks, IEEE Transactions on, 2006. **17**(4): p. 879-892.
38. Peng, H., F. Long, and C. Ding, *Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.* IEEE Transactions on Pattern

Analysis and Machine Intelligence, 2005. **27**(8): p. 1226-1238.

39. Cover, T.M. and J.A. Thomas, *Elements of information theory*. 1991, John Wiley.
40. Kira, K. and L.A. Rendell. *A practical approach to feature selection*. in *Proceedings of the ninth international workshop on Machine learning*. 1992. Morgan Kaufmann Publishers Inc.
41. Press, W.H., et al., *Numerical recipes in C: the art of scientific computing*. 1992, Cambridge Univ. Press, New York.
42. Little, M.A., et al., *Suitability of dysphonia measurements for telemonitoring of Parkinson' disease.* IEEE Transactions on Biomedical Engineering, 2009.
43. Witten, I.H., E. Frank, and M.A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques (third edition)* 2011, Burlington, MA: Morgan Kaufmann.
44. Fayyad, U. and K. Irani, *Multi-interval discretization of continuous-valued attributes for classification learning*, in *Proceedings of the 13th Int. Joint Conference on Artificial Intelligence*. 1993. p. 1022-1027.
45. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. 1995. C.S. Mellish (Ed.), Proceedings IJCAI-95, Montreal, Que., Morgan Kaufmann, Los Altos, CA (1995), 1137-1143.
46. Fawcett, T., *An introduction to ROC analysis.* Pattern recognition letters, 2006. **27**(8): p. 861-874.
47. Fawcett, T., *ROC graphs: Notes and practical considerations for researchers.* Machine Learning, 2004. **31**: p. 1-38.
48. Hsu, C.W., C.C. Chang, and C.J. Lin, *A practical guide to support vector classification*. 2003, Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003. available at *http://www.csie.ntu.edu.tw/cjlin/libsvm/*.

***Biography of the author(s)**
**Click here to download Biography of the author(s): bio.pdf**

Hui-Ling Chen is currently a lecture in the department of computer science and technology at Wenzhou University, China. He received his Ph.D. degree in department of computer science and technology at Jilin University, China. His present research interest centers on machine learning and data mining, as well as their applications such as medical diagnosis, bankruptcy prediction, gene selection, face recognition, among others. He is currently a Reviewer for IEEE Transactions on Systems, Man, and Cybernetics, Part B. He has published more than 50 papers in international journals and conference proceedings, including Pattern Recognition, Expert Systems with Applications, Knowledge-Based Systems, Soft Computing, PAKDD, and among others.

Gang Wang is currently a lecturer in the School of Computer Science and Technology, Jilin University. Currently he is working in the area of feature selection and its applications including spam filtering, gene selection and image retrieval. He is also associated with the development of parallel package for multi-core platform. His research interests include data mining, pattern recognition and parallel computing.

Chao Ma received the Master's degree from the College of Computer Science and Technology, Jilin University, China, in 2010. From September 2010, he is pursuing the Ph.D. degree at the College of Computer Science and Technology, Jilin University. His current research interests include data mining, classification analysis and neural network.

Zhen-Nao Cai is pursuing the Ph.D. degree in the school of computer at Northwestern Polytechnical University, China. He is now also a researcher at Wenzhou University, China. He received his Master's degree from school of software Engineering of Huazhong University of science and technology. His main research interests include machine learning, pattern recognition, and data mining.

Wen-Bin Liu is a professor of the Department of Physics and Electronic Information Engineering, Wenzhou University, China. He achieved his Ph.D. at the Department of Control Science and Engineering, Huazhong University of Science and Technology in 2004. Then he worked as a post Ph.D. researcher at the same group for two years. In 2007, he visited the Institute for Systems Biology for one year. In 2013, he visited Texas A&M University; his major interests include computational biology, data mining, and pattern recognition, DNA computing and evolutionary algorithms.

Su-Jing Wang received the Master's degree from the Software College of Jilin University, Changchun, China, in 2007. From September 2008, he is pursuing to the Ph.D. degree at the College of Computer Science and Technology of Jilin University. He has published more than 30 scientific papers. He is One of Ten Selectees of the Doctoral Consortium at International Joint Conference on Biometrics 2011. He was called as *Chinese Hawkin* by the Xinhua News Agency. His research was published in IEEE Transactions on Image Processing, Neurocomputing, etc. His current research interests include pattern recognition, computer vision and machine learning. He also reviews for several top journals, such as IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Neural Networks and Learning Systems. For details, please refer to his homepage http://sujingwang.name.

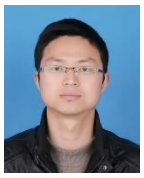
***Photo of the author(s)**

**Click here to download Photo of the author(s): photos.pdf**

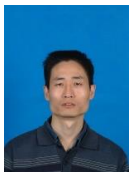
Hui-Ling Chen

Gang Wang

Chao Ma



Zhen-Nao Cai.



Wen-Bin Liu

Su-Jing Wang