Information Processing and Management 56 (2019) 609–623

Contents lists available at ScienceDirect

Information Processing and Management

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

A novel intelligent classification model for breast cancer diagnosis

Na Liu[a,b], Er-Shi Qi[a], Man Xu[c,*], Bo Gao[d], Gui-Qiu Liu[e]

[a] College of Management and Economics, Tianjin University, Tianjin, 300072, China
[b] School of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China
[c] Business School, Nankai University, Tianjin, 300071, China
[d] School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[e] Key Laboratory of Artificial Cell, Department of Pathology, The Third Central Hospital of Tianjin Medical University, Tianjin, 300170, China

ARTICLE INFO

Keywords:
Support vector machine
Breast cancer diagnosis
Information gain
Cost-sensitive learning
Genetic algorithm

ABSTRACT

Breast cancer is one of the leading causes of death among women worldwide. Accurate and early detection of breast cancer can ensure long-term surviving for the patients. However, traditional classification algorithms usually aim only to maximize the classification accuracy, failing to take into consideration the misclassification costs between different categories. Furthermore, the costs associated with missing a cancer case (false negative) are clearly much higher than those of mislabeling a benign one (false positive). To overcome this drawback and further improving the classification accuracy of the breast cancer diagnosis, in this work, a novel breast cancer intelligent diagnosis approach has been proposed, which employed information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection, in this process, we performs the ranking of features according to IG algorithm, and extracting the top *m* optimal feature utilized the cost sensitive support vector machine (CSSVM) learning algorithm. Our proposed feature selection approach which can not only help to reduce the complexity of SAGASW algorithm and effectively extracting the optimal feature subset to a certain extent, but it can also obtain the maximum classification accuracy and minimum misclassification cost. The efficacy of our proposed approach is tested on Wisconsin Original Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) breast cancer data sets, and the results demonstrate that our proposed hybrid algorithm outperforms other comparison methods. The main objective of this study was to apply our research in real clinical diagnostic system and thereby assist clinical physicians in making correct and effective decisions in the future. Moreover our proposed method could also be applied to other illness diagnosis.

1. Introduction

Breast cancer is one of the leading causes of death among women worldwide (Sheikhpour, Ghassemi, Yaghmaei, Ardekani, & Shiryazd, 2014). According to the American Cancer Society (ACS), an estimation of 246,660 women will be diagnosed with breast cancer and approximately 40,450 women will die from this disease in 2016 (American Cancer Society. 2016). In China, there has been an estimated of 214,360 women has died from breast cancer by 2008, and the number of death will reach up to 2.5 million by 2021 (Fan et al., 2014). However, according to the survey, more than 30% of cancer cases will be surviving for a long-time if they accept the accurate early detection (Sizilio, Leite, Guerreiro, & Neto, 2012). Thus, it is imperative for us to design an accurately and

[*] Correspondence author.

E-mail addresses: naliu188@tju.edu.cn (N. Liu), esqi@tju.edu.cn (E.-S. Qi), td_xuman@nankai.edu.cn (M. Xu), e17301115@stu.ahu.edu.cn (B. Gao), xasdguiqiu@gmail.com (G.-Q. Liu).

https://doi.org/10.1016/j.ipm.2018.10.014
Received 20 October 2017; Received in revised form 16 August 2018; Accepted 18 October 2018
Available online 09 January 2019

reliably classifier to detect the malignant tumors form benign ones at the early stage. To the best of our knowledge, the common methods for detecting breast cancer are mammography and fine needle aspiration cytology (FNAC), but these diagnostic techniques have demonstrated relatively low reliability for the detection of malignant tumors (Chen, Yang, Liu, & Liu, 2011). In recent years, with the development of artificial intelligence, more and more data-driven intelligent classification approaches have been applied for breast cancer diagnosis, such as Naïve Bayesian (Karabatak, 2015), Neural Network (Bhardwaj & Tiwari, 2015), Support Vector Machine (SVM) (Chen et al., 2011) or other hybrid algorithms (Ahn & Kim, 2009; Peng et al., 2016; Sun, Tseng, Zhang, & Qian, 2017; Gu et al, 2017; Qiu et al., 2017). But the fatal shortcoming of these excellent classification models is that they only pursuit of maximizing the classification accuracy, failing to consider the misclassification costs between different categories. To the best of our knowledge, the cost associate with missing a cancer case (false negative) is clearly much higher than those of mislabeling a benign one (false positive) (Krawczyk, Schaefer, & Woźniak, 2015). To make it clear, compare the cost of misclassifying a non-cancerous patient as cancerous to the cost of misclassifying a breast cancerous patient as non-cancerous, the former case may spend more cost associated with unnecessary biopsies for pathological analyses, but in the latter case, which may lead the patient missing the timely treatment and lead to death. Consequently, it is greatly important and imperative for the researchers to design an effective algorithm, which considering the unequal misclassification cost for the diagnosing of breast cancer. As for this, in this work we propose a novel breast cancer intelligent diagnosis method, which employs IGSAGAW for feature selection, and propagates the top *n* features through the CSSVM learning algorithm for breast tumor classification. In our proposed approach, we not only take into consideration the unequal misclassification costs of the breast cancer tumor, but also considering the average classification accuracy (ACC) of the breast cancer diagnosis.

The main advantage of feature selection in actual situation is that it can not only reduce the costs associated with unnecessary biopsies for pathological analyses, but also reduce the patients' waiting time without sacrificing the detection accuracy. In our work, we introduce IGSAGAW method. IG is a good measure to determine the relevance of feature and category (Jadhav et al., 2018; Soufan, Kleftogiannis, Kalnis, & Bajic, 2015), and moreover it often act as an importance measurement for the features. We firstly calculate the IG of each feature and rank the features according to its importance, then propagate the top *n* features through SAGAW algorithm. In this process, we utilize BP neural network, 3-NN and CSSVM classifiers as underlying classifiers. In our study, we utilize improved meta-heuristic search method of SAGA for feature selection. To the best of our knowledge, GA has the advantage of searching for the optimal solution quickly but it has a fatal shortcoming that is it is liable to be trapped in local optimal, due to this deficiency we introduce SA method to improve GA, which can effectively changing the annealing temperature during the iteration process and avoid trapping into local optimum (Dai, Tang, Giret, Salido, & Li, 2013; Li, Han, & Wang, 2015). Additionally, in the next step of our approach, we applied CSSVM classifier. In our work, we take both the average misclassification costs (AMC) of different classes and average classification accuracy (ACC) into account, and present an effective CSSVM classifier to accurately and reliably distinct malignant breast tumors from benign ones. Herein, due to SVM has been proved to be one of the most effective methods on addressing binary classification problem, and it has strong generalization performance and classification precision comparing with other conventional classification approaches (Akbani, Kwek, & Japkowicz, 2004; Chen et al., 2011; Sørensen & Nielsen, 2018), therefore in our study we utilized SVM to perform the classification task. Herein, we adopt RBF kernel function which has been considered widely utilized as in the SVM classification model (Chang & Lin, 2011). The main contributions of our work can be summarized as follows:

- Taking into consideration the advantages of IG and GAW, then proposing a IGSAGAW hybrid feature selection approach, which can remove redundant and irrelevant feature from the feature space, thus improving the classification accuracy and reducing the computational cost.
- Taking into consideration the unequal misclassification costs and classification accuracy (ACC), we propose a novel intelligent classification model to distinguish benign breast tumors from malignant ones. The effectiveness of our proposed method are verified on WBC and WDBC data sets, the empirical results demonstrate that our proposed method can achieve good performances.

The rest of this paper is organized as follows. Section 2 presents related works on breast cancer diagnosis. Section 3 presents the research objective of our study. Section 4 introduces the backgrounds and preliminaries of our method. Section 5 proposes the framework of our proposed approach. Section 6 presents the experimental analysis of our proposed model. Section 7 presents the discussion of our proposed model. Finally, Section 8 presents the conclusions of our research.

2. Related works

In this section, we summarize the previous studies of the breast cancer diagnosis over 10 years. Existing studies primarily adopt artificial neural networks (ANNs), decision tree analysis (DTA), Naïve Bayes (NB), Support Vector Machines (SVM) and so on. As Table 1 present the previous studies of breast cancer intelligent diagnosis in recent 10 years.

Due to the neural network has the advantages of capturing the correlations between attributes, therefore it has been widely utilized for breast cancer diagnosis (Lundin et al., 1999; Ravdin & Clark, 1992; Yao, 1999). Liu, Wang, and Zhang (2009) designed a decision tree prediction model for breast cancer survivability and adopt under-sampling method to balance the training data, the results has shown that when the ratio is equal to 15%, the AUC of the model is 0.7484. On top of the decision tree algorithm, Quinlan (1996) introduced MDL-inspired penalty and designed an improved C4.5 decision tree algorithm for breast cancer prediction, and attained the prediction accuracy of 94.74%. However, the performance of single learning classification algorithm can't reflect the

Table 1
Summary of typical previous research for breast cancer intelligent diagnosis.

Author	Year	Methods	Results
Akay	2009	SVM with F-score feature selection	Highest classification accuracy = 98.53%, 99.02%, 99.51% for 50% − 50%, 70–30%, 80–20% training-test partition
Ahn et al.	2009	Novel CBR	Acc = 99.12%
Chen et al.	2011	Rough set (RS) and SVM	Highest classification accuracy = 99.41%, 100%, 100% for 50% − 50%, 70–30%, 80–20% training-test partition
Zheng et al.	2014	K-means and SVM.	Acc = 97.38%
Onan	2015	Fuzzy-rough nearest neighbor	Acc = 99.7151%
Karabatak	2015	Naïve Bayesian(NB)	Sensitivity = 99.11%; Specificity = 98.25%; Accuracy = 98.54%
Sun et al.	2017	Deep convolutional neural network (DCNN)	AUC = 0.8818; Accuracy = 82.43%
Wang et al.	2018	SVM based ensemble learning algorithm	Accuracy = 97.89%

interactive factors of the breast cancer survival and recurrence rate (Wang, Zheng, Yoon, & Ko, 2018), Therefore, In order to overcome the drawbacks bring by single algorithm, numerous hybrid algorithms have been proposed. Akay (2009) presents *F*-score method for feature selection and SVM for breast cancer prediction. On top of that, another hybrid algorithm presented by Chen et al. (2011) which designed an hybrid classifier with rough set for feature selection and SVM for classification. Zheng, Yoon, and Lam (2014) proposed K-means and SVM hybrid algorithm for breast cancer diagnosis, K-means method for breast tumor feature extraction and SVM for classification, In another study, Onan (2015) designed a hybrid intelligent classification model for breast cancer diagnosis, which consist of fuzzy-rough approach for instance selection, consistency-based for feature selection and fuzzy-rough nearest neighbor algorithm for breast tumor classification. In addition, Sheikhpour, Sarram, and Sheikhpour (2016) proposed PSO and non-parametric kernel density estimation (KDE) based classifier to diagnose breast cancer. To summarize, their results shown that the proposed hybrid models have achieved high classification accuracy with fewer feature variables. In 2018, Wang et al designed an ensemble algorithm fusion SVM for breast cancer diagnosis which emphasis on model structures, and the results shown that the model achieves a higher accuracy compared to other ensemble models. To sum up, the main disadvantages of previous studies in breast cancer diagnosis is that they only pursuit of high classification accuracy, ignoring the unequal misclassification cost. Nevertheless, to the best of our knowledge, in medical diagnosis, comparing the cost of misclassifying a cancerous patient as a non-cancerous to misclassifying a non-cancerous patient as cancerous, the consequences may vary greatly. Therefore, this study constructs a hybrid intelligent classification model which has a competitive performance compared to other existing methods. The main advantage of our proposed classification model is that it can not only achieve the minimum misclassification cost, but also obtain the maximum classification accuracy with fewer input features.

3. Research objectives

In this work, we develop an efficient hybrid intelligent classification model for breast cancer diagnosis. There have two main advantages of our proposed model: the first is our study take fully account of the input feature dimensional, and design an effective feature selection method to select the optimum feature subset; the second is that this intelligent classification model can achieve the maximum classification accuracy of the breast cancer, and at the same time obtain the minimum misclassification cost. The main objectives of our proposed classification model are as follows:

1. Investigate the performance of IGSAGAW for feature selection. For this research objective, we compared our proposed method with GAW, and furthermore, in order to strengthen the significance of feature selection, we also carry out the comparative experiments with all features before applying feature selection approaches.
2. Examine the performance of CSSVM method. For this research objective, we carry out the comparative experiments with the same feature selection approaches, based on BP, 3-NN and CSSVM classifiers. And evaluate the performances of ACC, AMC, G-mean and running time.

4. Backgrounds and preliminaries

This section presents some preliminaries of our proposed method.

4.1. Information gain method

In this paper, we introduce IG directed SAGAW method for feature selection. To the best of our knowledge, the value of IG of each cases can represent its relevance to the category (Lai, Yeh, & Chang, 2016; Martín-Valdivia, Díaz-Galiano, Montejo-Raez, & Uren˜a-Lo´pez, 2008; Yang, Liu, Zhu, Liu, & Zhang, 2012), that is, a higher IG value means that the attribute contribute more information. For the classification system, we assume that the target dataset has $N = \{1,2,\ldots,n\}$ instances with k classes. Let $P(C_i, N)$ represents the proportion of C_i to N, where C_i represent the set of instances that belong to the ith class. The entropy of the dataset can be calculated by formula (1):

$$Entropy(N) = -\sum_{i=1}^{k} P(C_i, N) \times \log P(C_i, N) \tag{1}$$

If a case γ has $C = \{c_1, c_2, ...c_k\}$ distinct category and letting $N_i \in N|\gamma{=}c_i$, then the entropy of the dataset from category γ is given by:

$$Entropy_\gamma(N) = \sum_{i=1}^{n} \frac{|N_j|}{N} \times Entropy(N_j) \tag{2}$$

Finally, the value of IG of category γ can be derived by:

$$IG(\gamma) = entropy(N) - entropy_\gamma(N) \tag{3}$$

4.2. GA method

GA is a well-known global search method, which has received much attention for feature selection researchers (Dong, Li, Ding, & Sun, 2018; Ghosh, Parui, & Majumder, 2015; Hsu, 2004; Jadhav et al., 2018). GA can produce promising solutions for feature selection over a high-dimension space due to its robustness to the underlying search space size and multivariate distributions. To the best of our knowledge, the basic process of GA algorithm is as follows (Dong et al., 2018): (1) initialization. Random generate N individuals as the initial population, and encode the individuals; (2) individual evaluation. Calculate the fitness of each individual according to the evaluation criteria; (3) population evolution. Employ the selection operation, the crossover operation and the mutation operation to produce the next generation; (4) termination test. To judge if the maximum fitness of the individual is the optimal solution, if "yes", then terminate the calculation, otherwise return to (2).

4.3. SA method

SA is a heuristic global optimization method, which introduce Metropolis acceptance criteria to judge whether to accept a new solution or not (Javidrad et al., 2018; Liang, Suganthan, Chan, & Huang, 2006). The basic idea of SA is to start from an initial solution, and then integrate with the Metropolis Monte Carlo procedure. The first iterative process of SA is generate new solution, then judging whether it meet with Metropolis criterion, if "yes" then accept it, otherwise abandon it. The acceptance probability P of a candidate solution x_{i+1} from the current solution x_i is stated as:

$$P = \begin{cases} 1 \;\; iff(x_{i+1}) \le f(x_i) \\ e^{-\Delta f/T} otherwise \end{cases} \tag{4}$$

where f is the objective function, $\Delta f = f(x_{i+1}) - f(x_i)$ and T is a controlling parameter. As the solution procedure proceeds, temperature is sequentially lowered to reduce acceptance probability. The temperature decreasing rule has the form as below:

$$T_{i+1} = \xi T_i \tag{5}$$

where T_0 is the initial temperature and ξ is the temperature decay factor, which usually taken in the interval of 0–1.

4.4. SVM for classification

The SVM was originally developed by Vapnik (1995), and it is based on the Vapnik–Chervonenkis (VC) theory and structural risk minimization (SRM) principle (Vapnik, 1995, 1999). It can automatically determine the classification of data samples with a superior ability to distinguish the support vector, and the resulting classifier can maximize the interval between the class, leading to a better generalization ability and a higher classification accuracy. In addition, one major advantage of the SVM is the utilization of convex quadratic programming, which provides only global minimal and hence avoids trapping in local minimum (Vapnik, 1995; Cristianini & Shawe-Taylor, 2000). In our work, we take into consideration the unequal misclassification costs and introduce CSSVM method performing classification task. The more details about CSSVM can refer to Section 4.2. Herein, our proposed hybrid model has been implemented on the LIBSVM package (Chang & Lin, 2011) and employed Radial Basis Function (RBF) kernel.

5. The framework of our proposed approach

This section proposed the framework of our proposed approach. In our work, we firstly employs IGSAGAW for feature selection, which rely on IG ranking the importance of feature and help to reduce the computing complexity of SAGA wrapper approach, then we filtering some of the redundancy and unrelated features, and thereby extracting the top m optimal feature utilize the CSSVM learning algorithm. However, to apply this model to breast cancer diagnosis, there has an urgent problem need to be solved that is how to construct the fitness function. The selection of the fitness function is very important, which directly determine whether or not an optimal solution can be found. In this paper, we take fully account of the misclassification cost and the classification error rate, and construct the fitness function as follows:

Table 2
The cost matrix used by the classifiers.

True	Predicted	
	Benign/majority class	Malignant/minority class
Benign/majority class	0	$\cos t_{bm}$
Malignant/minority class	$\cos t_{mb}$	0

$$F = \frac{(f_{mb} \cos t_{mb} + f_{bm} \cos t_{bm})^{*}(f_{mb} + f_{bm})}{n^2} \tag{6}$$

In formula (6), f_{mb} and $cost_{mb}$ are the number of samples and the misclassification cost of the malignant tumors diagnosed as benign ones; f_{bm} and $cost_{bm}$ are the number of samples and the value of the benign tumors diagnosed as malignant ones and n represents the number of samples. The cost matrix is shown in Table 2. where the $cost_{bm}$ is the misclassification cost associated with the benign tumor assigned to the category of malignant tumor, and $cost_{mb}$ is the misclassification cost associated with the malignant tumor assigned to the category of benign tumor.

5.1. The approach of CSSVM for classification

In cost-sensitive decision system. We define D as a decision variable, which including z th category label of $\{d(1), d(2), \ldots, d(z)\}$. As for each input and output variable, a sample pair can be formed as follows:

$$\{\mathbf{a}(\mathbf{i}), d(i), cost(i)\}, \quad i = 1, 2, \ldots, n \tag{7}$$

Where in formula (7), $\mathbf{a(i)} = \{a_1(i), a_2(i), \ldots, a_m(i)\}$ represent the input samples, and m is the number of features, n is the number of cases. $d(i)$ represent the decision category, and there have z decision category $d(k) = \{d(1), d(2), \ldots, d(z)\}$. For binary-classification problems, we define $d(k) = \{d(1), d(2)\}$. and in our work, we assumed that the negative category of $d(1) = -1$, and the positive category of $d(2) = +1$. Additionally, in formula (7), $\cos t(i) \geq 0$ which represent the misclassification cost of different category.

This study integrated different misclassification costs information into the CSSVM classifier, and design the objective function based on the empirical cost and structural cost minimization. According to Xu, Zhou, and Chen (2018), the standard cost-insensitive SVM has shown in formula (8). where in formula (8), $(C_1 \sum_{i:y_i=+1} \xi_i + C_2 \sum_{i:y_i=-1} \xi_i)$ can fully represent the misclassification cost of false negative and false positive respectively.

$$\min_{w,b,\xi_i} \frac{1}{2} ||w||^2 + C_1 \sum_{i:y_i=+1} \xi_i + C_2 \sum_{i:y_i=-1} \xi_i$$
$$s.\ t. \begin{cases} y_i((k(w, \mathbf{x}_i) + b) \geq 1 - \xi_i; & i = 1, 2, \ldots, n \\ \xi_i \geq 0; & i = 1, 2, \ldots, n \\ C_i \geq 0; & i = 1, 2 \end{cases} \tag{8}$$

where in formula (8), C_1, C_2 represent the different penalty parameters of the breast tumor. ξ_i is a slack factor. And b is the threshold for the SVM decision boundary.

5.2. The main steps of our proposed approach

In order to comprehensively evaluate the experimental performance of our proposed approach, we present the main steps of our proposed hybrid intelligent diagnosis algorithm as shown in Fig. 1. The main steps of our proposed IGSAGAW algorithm are described below:

Step 1: Calculate the IG of the individual from the data set and rank the feature according to the value of the importance.
Step 2: Setting the initial parameters of GA and SA algorithms, then calculate the initial fitness according to formula (6).
Step 3: Set the initial generation and implement the GA iterative update steps.
Step 4: Calculate the fitness for each new individual according to CSSVM algorithm.
Step 5: Replace the least fitness population with new best individual.
Step 6: While gen < maxgen?, If "yes", then go to ***Step 4***, otherwise judge the current annealing temperature whether to meet with the end of the setting annealing temperature, if "yes", then output the optimal feature subset, otherwise implement the temperature decay operation of $T_{j+1} = \xi \times T_j$, and then return to ***Step3.***

The overall flow chart of IGSAGAW algorithm is shown in Fig. 1.

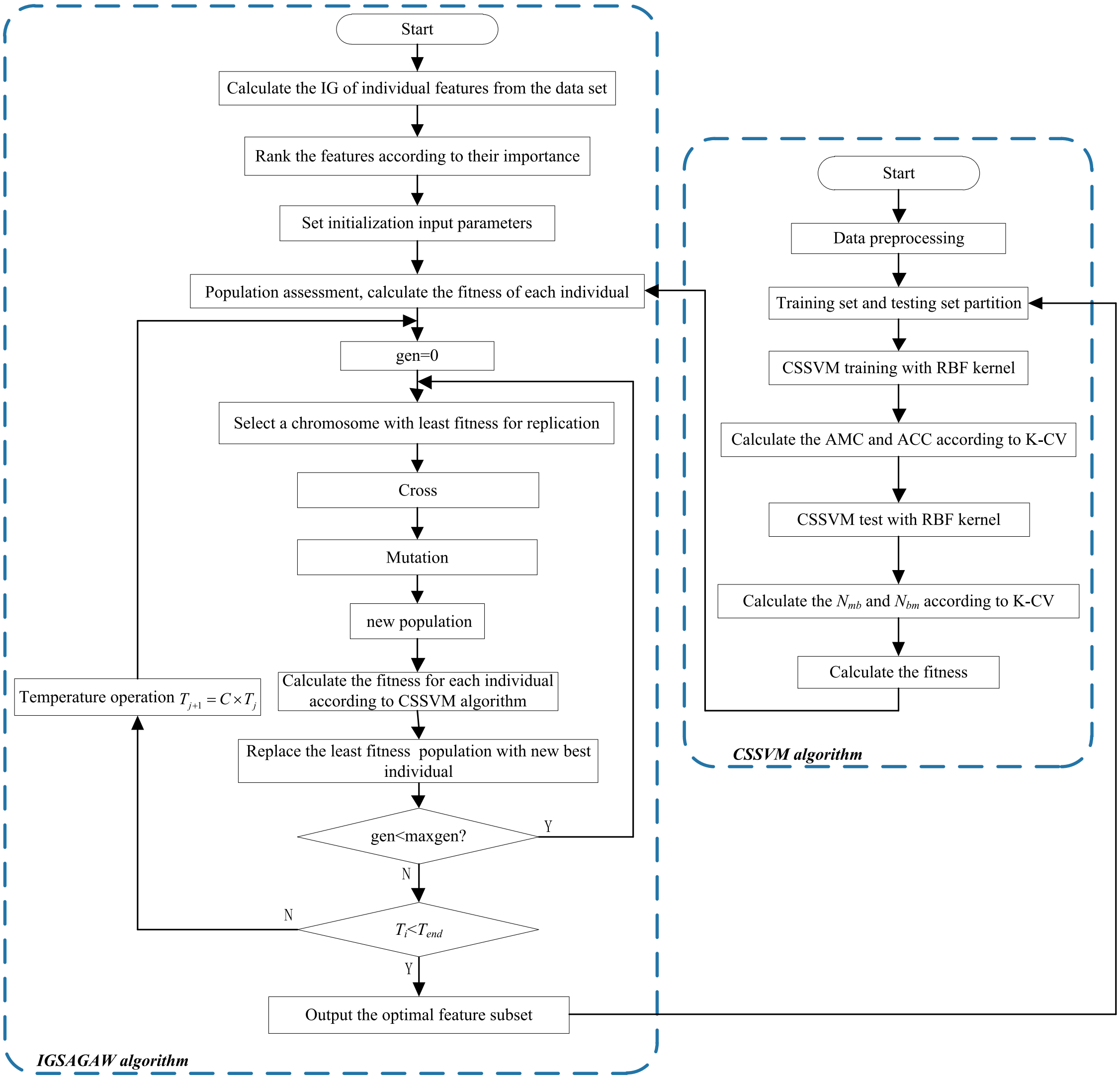
Start
Calculate the IG of individual features from the data set
Rank the features according to their importance
Set initialization input parameters
Population assessment, calculate the fitness of each individual
gen=0
Select a chromosome with least fitness for replication
Cross
Mutation
new population
Calculate the fitness for each individual according to CSSVM algorithm
Temperature operation $T_{j+1}=C\times T_j$
Replace the least fitness population with new best individual
gen<maxgen?
Y
N
$T_i<T_{end}$
N
Y
Output the optimal feature subset
IGSAGAW algorithm
Start
Data preprocessing
Training set and testing set partition
CSSVM training with RBF kernel
Calculate the AMC and ACC according to K-CV
CSSVM test with RBF kernel
Calculate the N_{mb} and N_{bm} according to K-CV
Calculate the fitness
CSSVM algorithm

Fig. 1. Framework of our proposed methods.

6. Experimental analysis

In order to examine the effectiveness and rationality of our proposed approach, we present the experimental setup in terms of the following three aspects: (1) the data sets used in the experiments, (2) the evaluation measures of the experiments, and (3) the details about the implementation of the experiments. Our experimental analysis was carried out on the Matlab 2016 mathematical development environment and the performance parameters of the executing host were win 10, Inter (R) Core (TM) i5-82350U CPU, 1.80 Ghz, X64, and 16 GB (RAM). Herein we employ the libsvm toolbox developed by Chang and Lin (2011).

6.1. Data sets

To evaluate the performance of the proposed methods, the experiment were conducted based on WBC and WDBC data sets from the UCI repository (UCI Machine Learning Repository: Data sets), the details of the two data sets are presented in Table 3, and the

Table 3
Details of the two data sets.

Data set	Number of attribute	Number of cases	Class distribution(B/M)	Missing value
WDBC	32	569	357/212	0
WBC	10	699	458/241	16

Table 4
Summary of attributes for WBC data set.

Attribute	Domain	Mean	Standard error
Clump thickness	1–10	4.44	2.82
Uniformity of cell size	1–10	3.15	3.07
Uniformity of cell shape	1–10	3.22	2.99
Marginal adhesion	1–10	2.83	2.86
Single epithelial cell size	1–10	3.23	2.22
Bare nuclei	1–10	3.54	3.64
Bland chromatin	1–10	3.45	2.45
Normal nucleoli	1–10	2.87	3.05
Mitoses	1–10	1.60	1.73

class distribution of B and M represent benign tumor and malignant ones respectively. Additionally, the details of attribute information are presented in Tables 4 and 5.

6.2. Evaluation measures

To evaluate the performance of our proposed hybrid algorithm, the classification accuracy (AC), average misclassification cost (AMC) and G-mean are utilized as the evaluation approaches. To the best of our knowledge, G-mean which is the geometric mean of true positive rate (TPR) and true negative rate (TNR), are proposed to evaluate the performance of the classifier on imbalanced data (Piri, Delen, & Liu, 2018). The calculation formulas are presented as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

$$AMC = \frac{Number_{fp} \times \cos t_{mb} + Number_{fn} \times \cos t_{bm}}{TP + FN + FP + TN} \tag{10}$$

$$Sensitivity\,(TPR) = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity\,(TNR) = \frac{TN}{FP + TN} \tag{12}$$

$$G - mean = \sqrt{TPR*TNR} \tag{13}$$

The calculation of TPR and TNR are based on the confusion matrix, which is presented in Table 6. As it can be observed in Table 6, *TP* is the number of true positives, which represents cases that are correctly categorized in the benign tumor. *FN* is the number of false negatives, which represents as benign tumor cases that are misclassified as malignant ones; *TN* is the number of true negatives, which represents cases that are correctly categorized in the malignant tumor; and *FP* is the number of false positives, which represents malignant tumor cases that are misclassified as benign ones.

6.3. Experimental design

This section presents the details of our experimental design. In order to verify the effective performances of our proposed model, we design the corresponding comparative experiment from two aspects. In the first aspect, we will compare the performances of our proposed IGSAGAW feature selection approach with GAW method, as can be seen from Table 7, it presents the main parameters of SAGAW algorithm. Additionally, in order to emphasis the superiority performances of feature selection, we ran the experiment on the

Table 5
Rang of each attributes in WDBC data set.

Attribute	Mean	Standard error	Maximum
Radius	6.98–28.11	0.112–2.873	7.93–36.04
Texture	9.71–39.28	0.36–4.89	12.02–49.54
Perimeter	43.79–188.50	0.76–21.98	50.41–251.20
Area	143.50–2501.00	6.80–542.20	185.20–4354.00
Smoothness	0.053–0.163	0.002–0.031	0.071–0.223
Compactness	0.019–0.345	0.002–0.135	0.027–1.058
Concavity	0.000–0.427	0.000–0.396	0.000–1.252
Concavity points	0.000–0.201	0.000–0.053	0.000–0.291
Symmetry	0.106–0.304	0.008–0.079	0.157–0.664
Fractal dimensional	0.050–0.097	0.001–0.030	0.055–0.208

Table 6
Confusion matrix.

	Predicted positive (benign)	Predicted negative (malignant)
Actual positive (benign)	True positive(*TP*)	False negative(*FN*)
Actual negative (malignant)	False positive(*FP*)	True negative (*TN*)

Table 7
The main parameters of SAGAW algorithm.

Parameter	Value
Maximum number of generation	200
The size of the population	50
Selection type	Tournament selection
Cross type	Single point cross
Mutation Rate	0.1
Mutation Type	Uniform mutation
Initial temperature value	100
Temperature decay coefficient	0.9

basis of underlying classifiers with all features, then applying different feature selection approaches of GAW and IGSAGAW using three typical classifiers.

The aim of our second aspect is to explore the effect of the different underlying classifiers. To the best of our knowledge, in the field of machine learning, breast cancer diagnosis has been considered as classification problem, and the classification approaches such as BP neural network, K-NN, CSSVM were considered as excellent classifiers, which have been utilized as underlying classifiers in our experimental. And the parameters of different comparative methods are presented in Table 8. In this work, we adopt grid search approach for finding the optimal parameters of SVM and set the search range of parameter $C = \{2^{-10}, 2^{-8}, \ldots, 2^{8}, 2^{10}\}$ and $g = \{2^{-10}, 2^{-8}, \ldots, 2^{8}, 2^{10}\}$, respectively, then the value of *C, g* searching step is set to 0.5. The main objective of parameter searching is to find the best parameter pair of (*C, g*), which can achieve the best performances. After obtaining the best parameter pair, we create the classification model performing for training and testing. To ensure the rationality of our experimental results, we utilized the data set which processed by the feature selection method as described above, and adopt 10-fold cross validation for WBC and WDBC data sets. In each comparative approach, we take into account the different feature selection methods as described above and design the source codes, which were implemented in MATLAB platform. In order to eliminate the randomness factor and reflect the results rationality, we employed 10-fold cross verification and the average performance of these results were reported as the final results in this study.

6.4. Experimental results and analysis

In our experimental, we design the source code of our proposed algorithm, which implemented on MATLAB platform. The results of 10-fold cross validation on different performances as can be observed in Figs. 2–9. As previous mentioned, in order to straightforward assess the effectiveness of our proposed model, we carried out a series of comparative experiments, and the experimental results analysis as described below.

6.4.1. Classification accuracy for the best solution

First of all, in order to verify the effect of our proposed model, we first ran the experiments on the baseline classifiers with all the features before applying our IGSAGAW and GAW feature selection approaches. As can be seen from Figs. 2 and 3, the average classification accuracy of 10-fold cross verification on IGSAGAW is higher than GAW, and the accuracy of GAW is higher than baseline classifier. The main reason is that in GAW feature selection approach, we select the top *n* features which obtained the best value of fitness. And during this feature selection process, we applied BP, 3-NN and CS-SVM three underlying classifiers perform for classification. In IGSAGAW approach, we utilized IG ranking the importance of features firstly, then we applied SAGAW algorithm to

Table 8
The parameters of different comparative methods.

Comparative methods	Parameters
3-NN	Set the parameter of K equal to 3
CSSVM	The best parameter pair of (C, g) by means of grid search method, set the *C* range of variation $[2^{-10}, 2^{10}]$, *g* range of $[2^{-10}, 2^{10}]$, and step value of *C* and *g* is set to 0.5.
BP neural network	Set the initial epochs equal to 1000; The learning rate of BP is set to 0.1; The final goal of BP is 0.1;

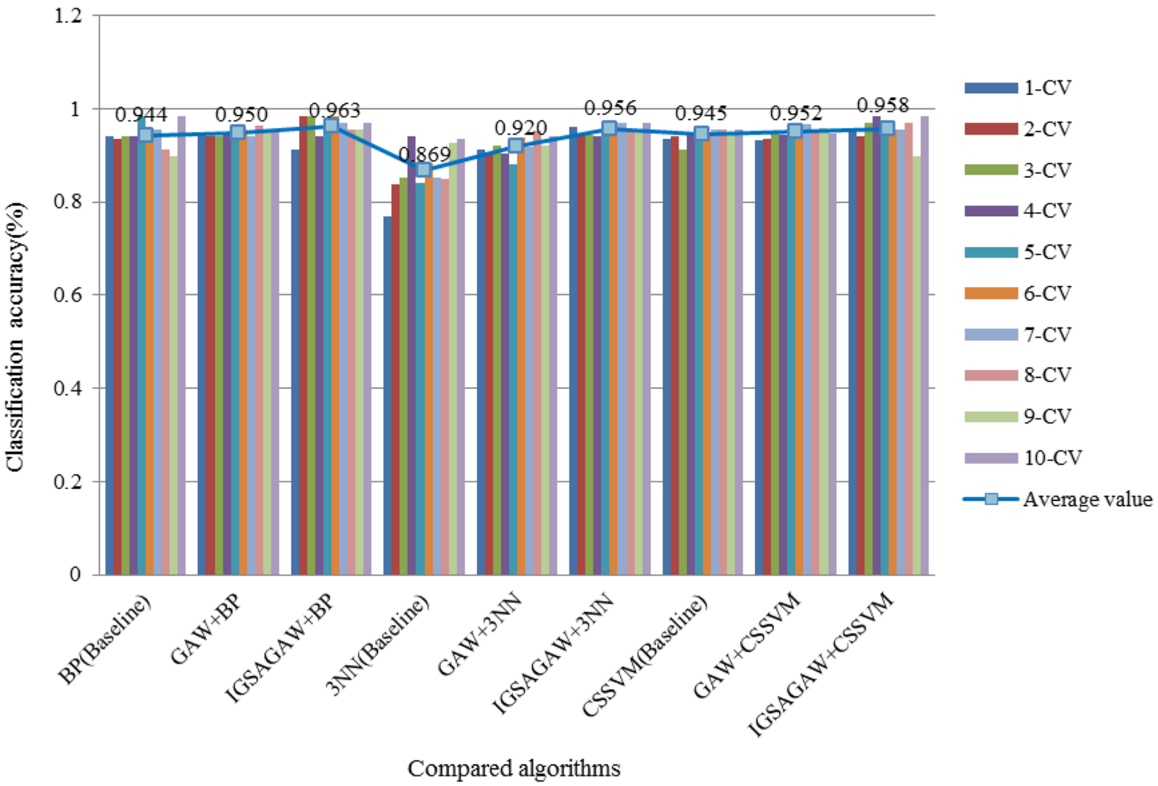

Fig. 2. Classification accuracy for different comparative methods based on WBC data set.

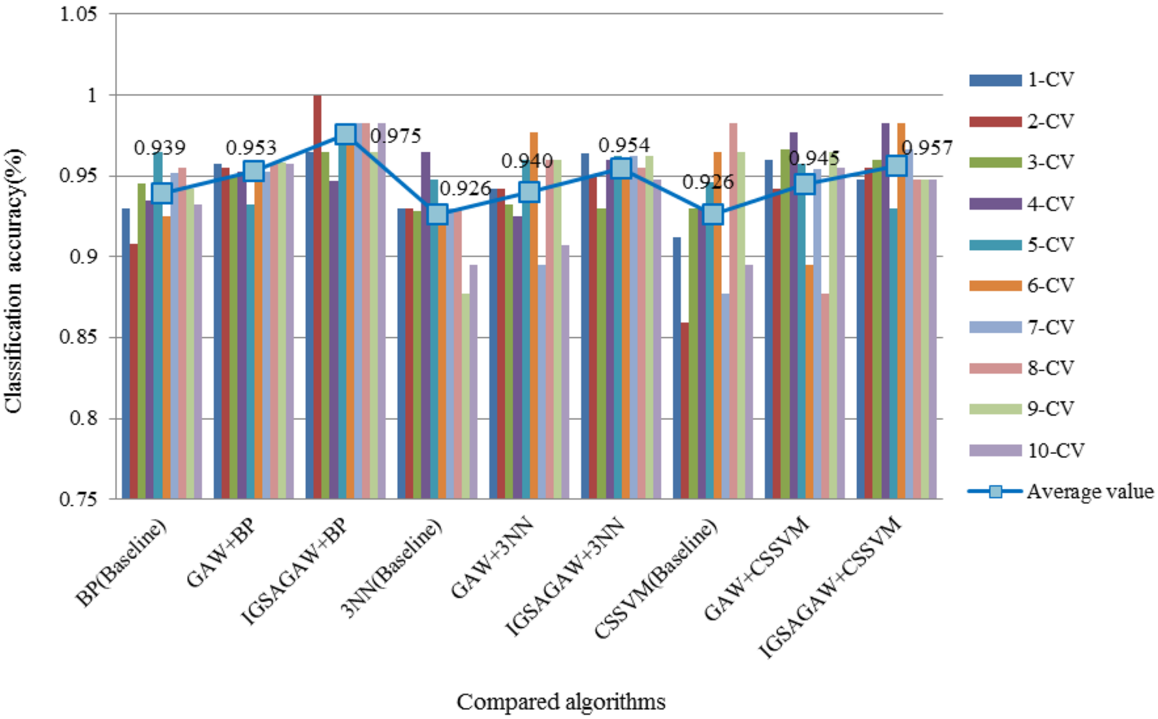

Fig. 3. Classification accuracy for different comparative methods based on WDBC data set.

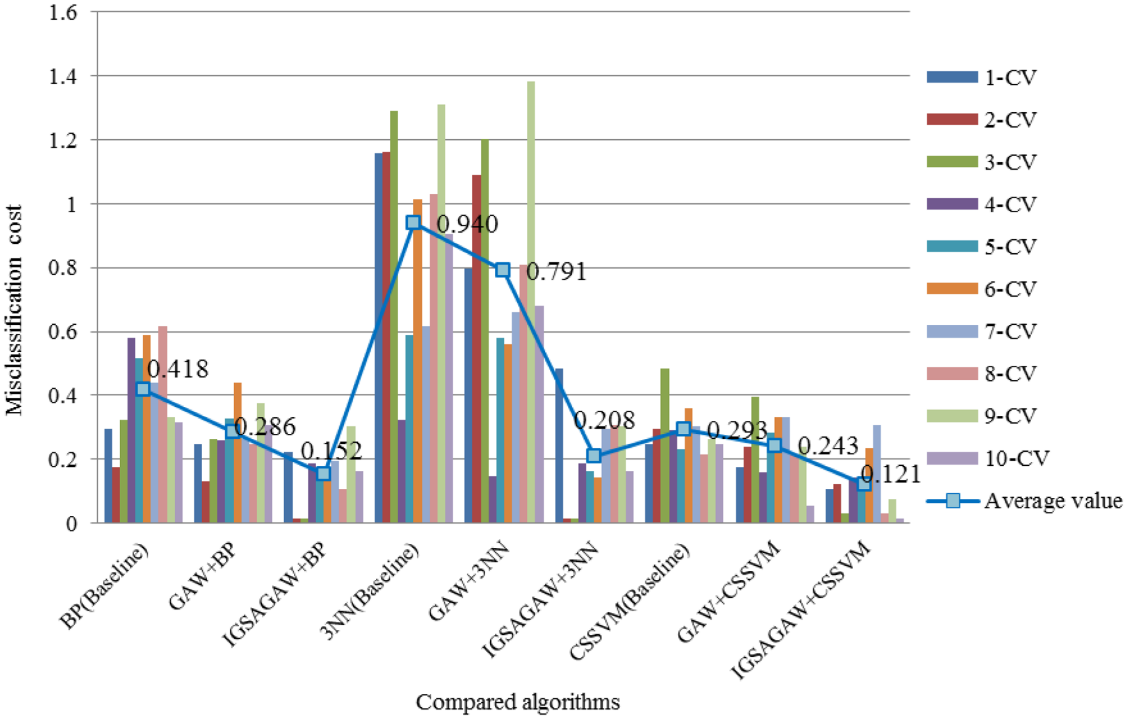

Fig. 4. Misclassification cost for different comparative methods based on WBC data set.

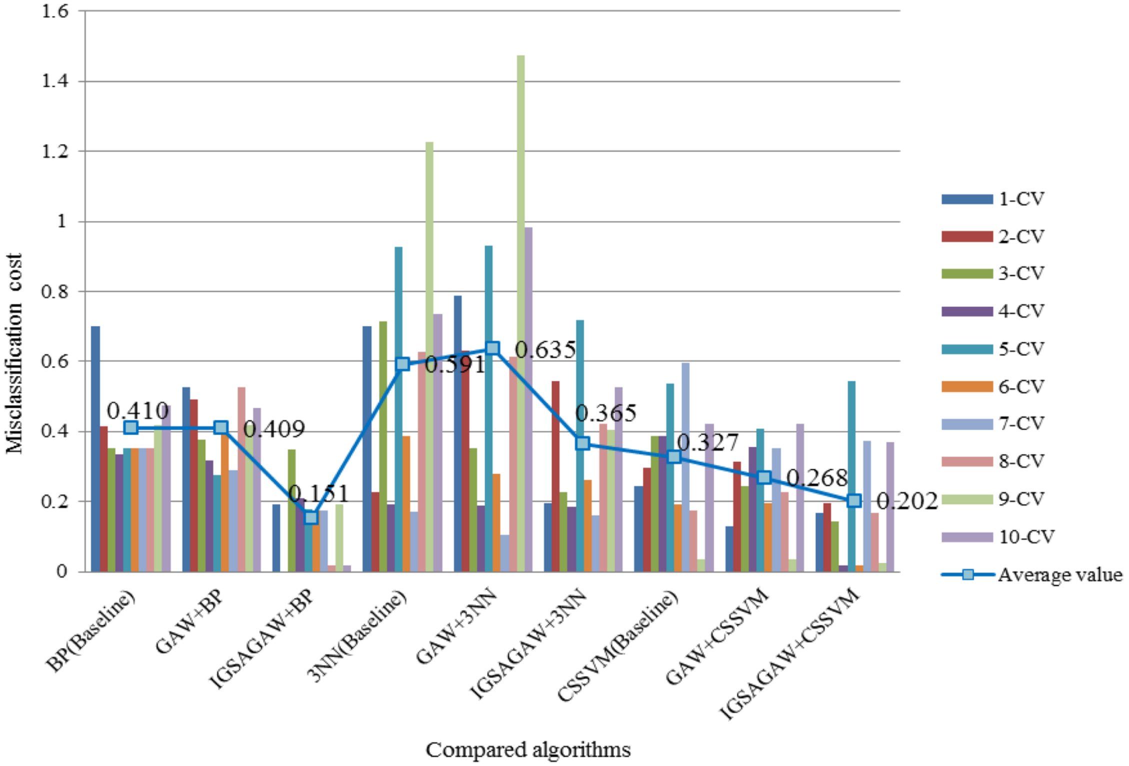

Fig. 5. Misclassification cost for different comparative methods based on WDBC data set.

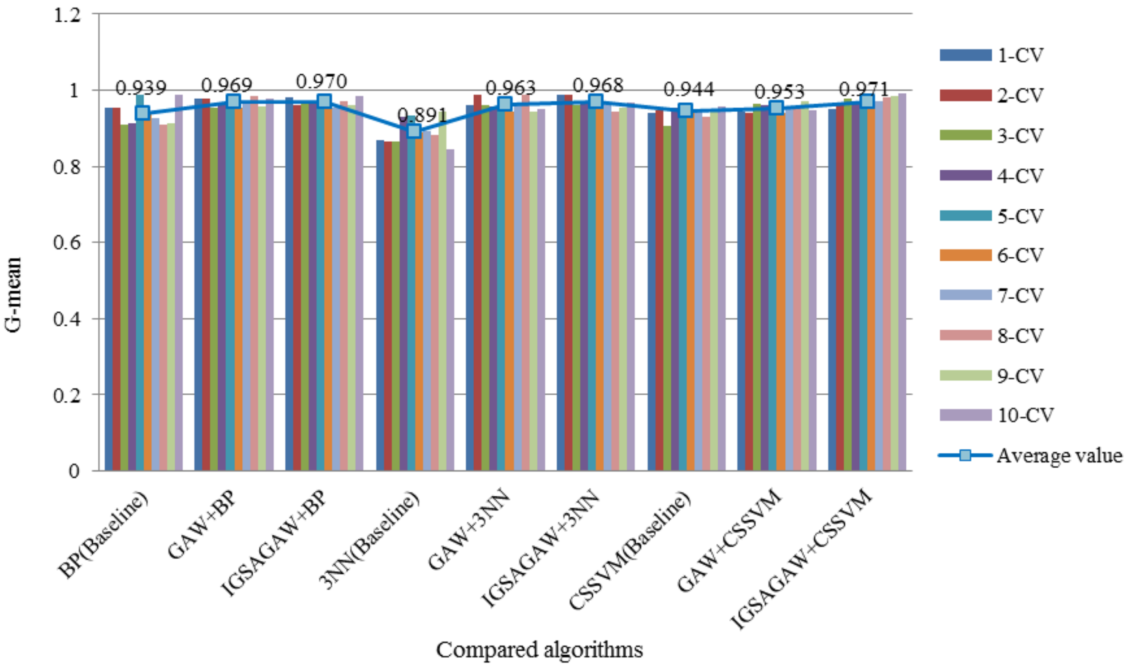

Fig. 6. G-mean for the different comparative methods based on WBC data set.

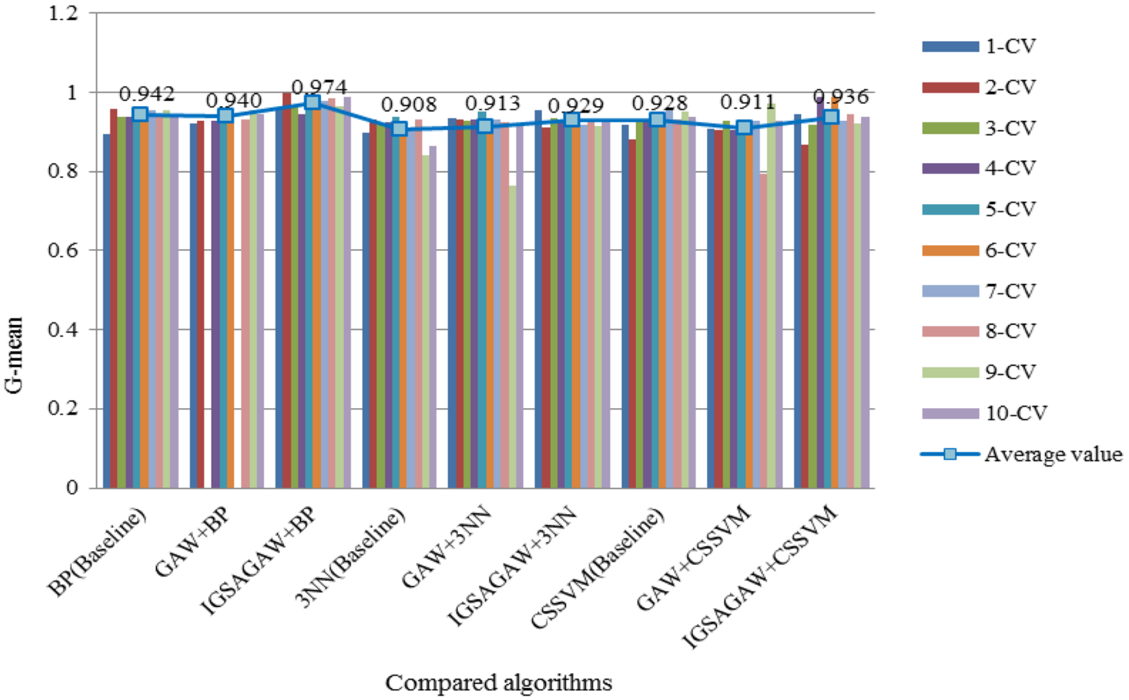

Fig. 7. G-mean for the different comparative methods based on WDBC data set.

select the top n features which obtained the best fitness. In this process, we adopt SA algorithm to optimize GA, which can avoid trapping into local optimum. Based on this, we can obtain the best results as presented in Figs. 2 and 3, which can be seen that the classification accuracies of SAGAW feature selection approach can achieve better performances than GAW methods.

6.4.2. *Misclassification cost for the best solution*

In our proposed model, we take fully account of misclassification cost of breast cancer tumor and quantification the

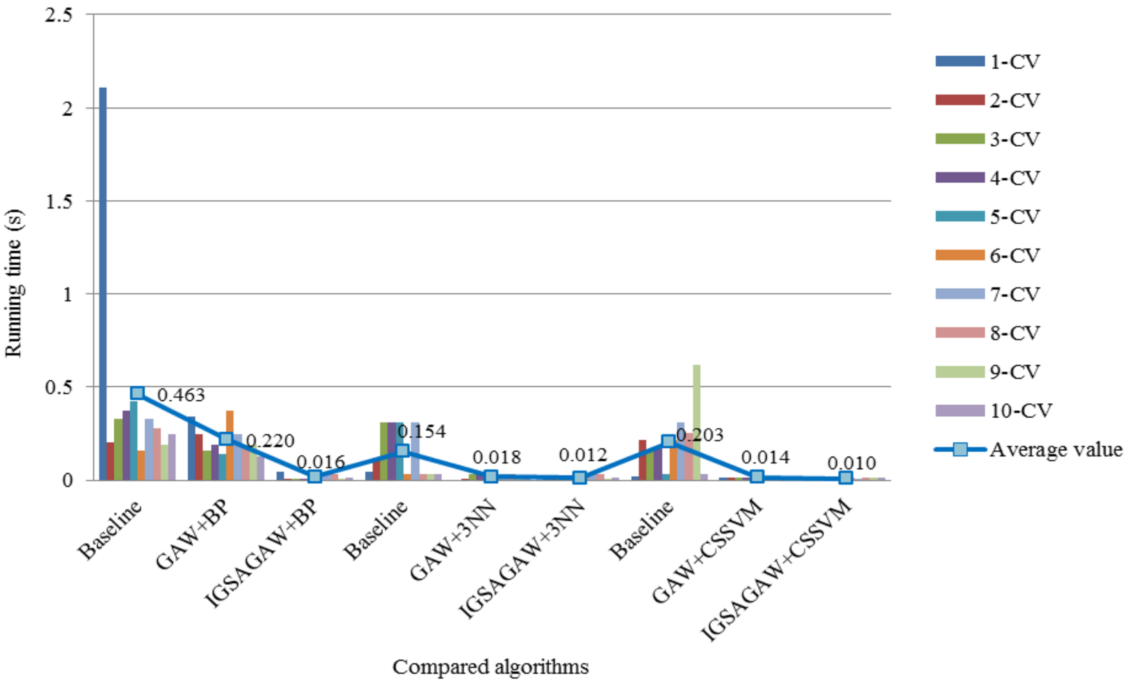

Fig. 8. Running time for the different comparative methods based on WBC data set.

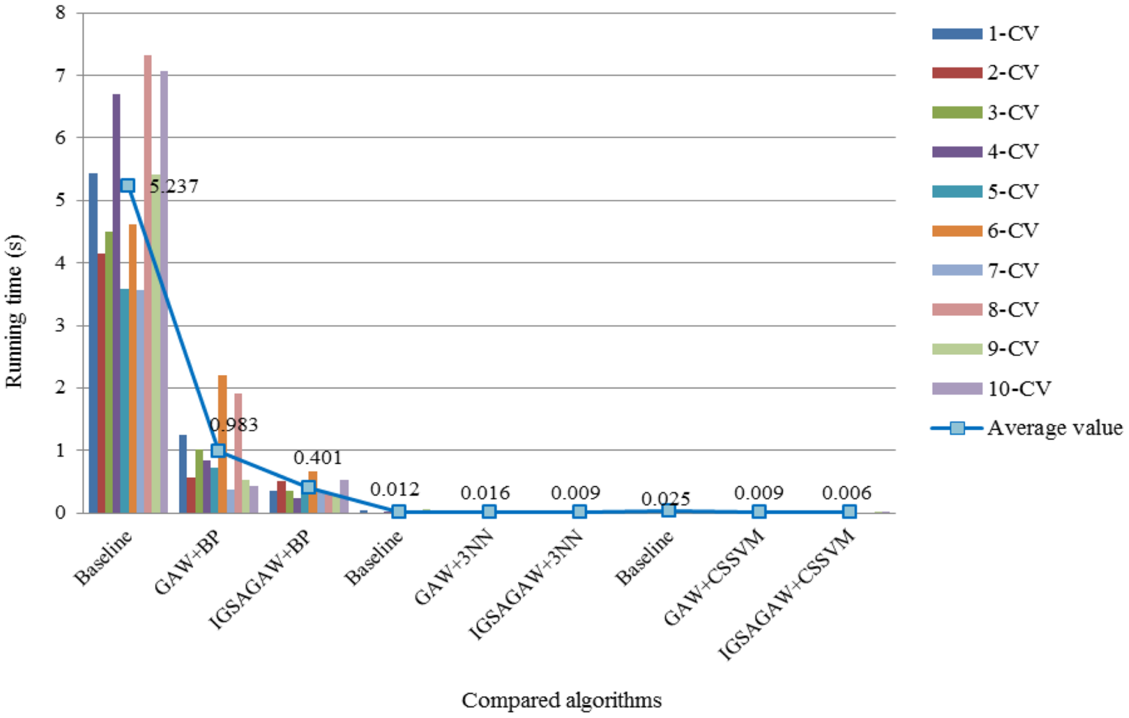

Fig. 9. Running time for the different comparative methods based on WDBC data set.

misclassification cost of two different scenarios as described above. In this work, we set the value of the correct classification cost as 0, and the misclassification cost had to further considering two scenarios: the first is misclassified malignant tumors as benign ones and the second aspect is misclassified benign tumors as malignant ones. As noted before, the consequences of this two scenarios vary greatly, herein in our work, we take fully account of expert experience and set $mc_{mb} = 10$ and $mc_{bm} = 1$ so as to make this two scenarios difference.

Table 9
The results of 10-fold cross verification based on WDBC data set.

Underlying classifier		Accuracy	Misclassification cost	G-mean	Running time
BP	Baseline	0.939	0.410	0.942	5.237
	GAW + BP	0.953	0.409	0.940	0.983
	IGSAGAW + BP	***0.975***[a]	***0.151***[a]	***0.974***[a]	0.401
3-NN	Baseline	0.926	0.591	0.908	0.012
	GAW + 3NN	0.940	0.635	0.913	0.016
	IGSAGAW + 3NN	0.954	0.365	0.929	0.009
SVM	Baseline	0.926	0.327	0.928	0.025
	GAW + CSSVM	0.945	0.268	0.911	0.009
	IGSAGAW + CSSVM	***0.957***[b]	***0.202***[b]	***0.936***[b]	***0.006***[b]

Note:
[a] Denotes the best results, but is not the optimum results.
[b] Denotes the optimum results.

The results of misclassification cost for different comparative methods are presented in Figs. 4 and 5, form the results we can obviously see that the misclassification cost of IGSAGAW algorithm achieved the best results followed by GAW algorithm, followed by the baseline approaches. The main reason is that in GAW feature selection approach, we select the top *n* features which can obtain the maximum classification accuracy and minimum misclassification cost. And during this feature selection process, we applied BP, 3-NN and CS-SVM three underlying classifiers perform for classification. In IGSAGAW approach, we utilized IG ranking the importance of features firstly, then we applied SAGAW algorithm to select the top *n* features which obtained the best fitness. In this process, we adopt SA algorithm to optimize GA, which can avoid trapping into local optimum and achieved the optimum results.

6.4.3. G-mean for the best solution

G-mean is the geometric mean of true positive rate (TPR) and true negative rate (TNR), which can be calculate by formulas (11) and (12). From the results of Figs. 6 and 7, we can obviously see that the IGSAGAW approaches achieved the best results followed by GAW approaches, followed by baseline approaches. The main reason is that the number of *FN* and *FP* in IGSAGAW is less than GAW, and followed by underlying approaches.

6.4.4. Running time for the best solution

Feature selection approaches can decrease the dimensional of feature space and reduce the computational complexity. Furthermore, it can reduce the running time for the classification models. As can be observed in Figs. 8 and 9, we can obviously see that the IGSAGAW approaches achieved the best results followed by GAW approaches, followed by the baseline approaches. The main reason is that in IGSAGAW approaches, we firstly ranking the features according to its importance, which can increase the computational efficiency and decrease the difficulty for searching.

From Figs. 2–9, it is seen that the IGSAGAW and GAW approaches have performed better than the baseline methods. The results of 10-fold cross validation on different models for WBC and WDBC data sets are concluded in Tables 9 and 10. The best and the optimum results are printed in bold. As can be observed in Tables 9 and 10, the results obviously indicate that our proposed model can achieve better performances than other comparative models.

7. Discussion

In this section, we will provide a discussion on the performance of different components of our proposed model. The proposed model is a hybrid intelligent classification method which fusion feature selection and classification. As previous mentioned, in order

Table 10
The results of 10-fold cross verification based on WBC data set.

Underlying classifier		Accuracy	Misclassification cost	G-mean	Running time
BP	Baseline	0.944	0.418	0.939	0.463
	GAW + BP	0.950	0.286	0.969	0.220
	IGSAGAW + BP	***0.963***[a]	0.152	0.970	0.016
3-NN	Baseline	0.869	0.940	0.891	0.154
	GAW + 3NN	0.920	0.791	0.963	0.018
	IGSAGAW + 3NN	0.956	0.208	0.968	0.012
SVM	Baseline	0.945	0.293	0.944	0.203
	GAW + CSSVM	0.950	0.243	0.953	0.014
	IGSAGAW + CSSVM	***0.958***[b]	***0.121***[b]	***0.971***[b]	***0.010***[b]

Note:
[a] Represents the best results, but is not the optimum results.
[b] Represents the optimum results.

to straightforward assess the effectiveness of our proposal, we carried out a series of comparative experiments.

First of all, in order to verify the effect of our proposed feature selection approach, we first ran experiments on baseline classifiers with all features before applying GAW and IGSAGAW approaches. And the results can be seen in Figs. 2–9, from the results we can indicate that the IGSAGAW approach achieved the best results followed by GAW method, followed by baseline classifiers. To strengthen the advantageous of feature selection methods, we also investigate the CPU running time of our proposed model, the results as presented in Figs. 8 and 9, from this two figures, we can clearly see that our proposed model can decrease the computational efficiency for the breast cancer diagnosis.

Moreover, in order to verify the efficiency of the CSSVM classification algorithm proposed in this paper, we compare it with BP neural network and 3-NN. The experimental result are shown in Figs. 8 and 9, From the result shown in Fig. 2, the classification accuracy of IGSAGAW + BP neural network model achieved the best results for the WDBC and WBC data sets, which are 97.5%, 96.3%, followed by IGSAGAW + CSSVM model, which are 95.7%, 95.8%, the details are presented in Tables 9 and 10, where in Tables 9 and 10, we can clearly see that the BP neural network approaches achieved the best results, which we marked in bold. At the same time, from Fig. 5 we can clearly see that the hybrid model of IGSAGAW + BP also obtained the best results in terms of misclassification cost for the WDBC data set, which is 0.151, followed by IGSAGAW + CSSVM model, which is 0.202. Moreover, as for the performance of G-mean, the result of hybrid model of IGSAGAW + BP is 0.974, which is better than our proposed is 0.936 for the WDBC data set. From the above analysis, it can be seen that the model of IGSAGAW + BP can achieve better results than our proposed model, which in terms of the performances of classification accuracy, misclassification cost and G-mean. However, the most important is that the running time of our proposed model is the minimum, which is far less than IGSAGAW + BP hybrid model. Therefore from the above analysis, it can be seen that our proposed model is the most suitable method; it produces an excellent performances and only requires a moderate computational cost for solving breast cancer classification problems.

8. Conclusion

In this research, a novel intelligent classification model has proposed for breast cancer intelligent diagnosis. This model firstly employed IG for feature ranking, then according to the importance ranking, this study introduced SAGAW hybrid method for feature selection, and propagates the top *m* features by means of CSSVM classifier for breast cancer diagnosis. All the results are conducted based on two standard breast cancer data set that are WBC and WDBC data sets. The results have shown that the proposed intelligent classification model not only can effectively evaluate the misclassification cost but also increase the breast cancer diagnosis performance and decrease the calculation complexity. To evaluate the performance of our proposed method, six comparison models have been introduced, and the empirical results indicated that our proposed ensemble learning method can achieve superiority performances than other comparison methods. The main objective of this work is to apply our research in real clinical breast cancer diagnostic system and thereby assist clinical physicians in making correct and effective decisions in the future.

Acknowledgments

We are grateful to the editor and the anonymous reviewers for their valuable comments and suggestions which have immensely improved the content of this version. This work was supported by the National Natural Science Foundation of China (71571105) and The Fundamental Research Funds for the Central Universities.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2018.10.014.

References

Ahn, H., & Kim, K. (2009). Global optimization of case-based reasoning for breast cytology diagnosis. *Expert Systems with Applications, 36*(1), 724–734.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications, 36*(2), 3240–3247.

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *European conference on machine learning* (pp. 39–50). .

American Cancer Society. (2016). *Cancer facts and figures 2016.* Atlanta: American Cancer Society.

Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. *Expert Systems with Applications, 42*(10), 4611–4620.

Chang, C. C., & Lin, C. J. (2011). *LIBSVM: A library for support vector machines.* ACM.

Chen, H., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications, 38*(7), 9014–9022.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines.* Cambridge, UK: Cambridge University Press.

Dai, M., Tang, D., Giret, A., Salido, M., & Li, W. (2013). Energy-efficient scheduling for a flexible flow shop using an improved genetic-simulated annealing algorithm. *Robotics and Computer-Integrated Manufacturing, 29*(5), 418–429.

Dong, H., Li, T., Ding, R., & Sun, J. (2018). A novel hybrid genetic algorithm with granular information for feature selection and optimization. *Applied Soft Computing, 65*, 33–46.

Fan, L., Strasserweippl, K., Li, J. J., et al. (2014). Breast cancer in China. *The Lancet Oncology, 15*(7), e279–e289.

Ghosh, K., Parui, S. K., & Majumder, P. (2015). Learning combination weights in data fusion using genetic algorithms. *Information Processing & Management, 51*(3), 306–328.

Gu, D., Liang, C., & Zhao, H. (2017). A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis. *Artificial Intelligence in Medicine, 77*, 31–47.

Hsu, W. H. (2004). Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning. *Information Sciences, 163*(1-3), 103–122.

Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing, 69*, 541–553.

Javidrad, F., Nazari, M., & Javidrad, H. R. (2018). Optimum stacking sequence design of laminates using a hybrid PSO-SA method. *Composite Structures, 185*, 607–618.

Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. *Measurement, 72*, 32–36.

Krawczyk, B., Schaefer, G., & Woźniak, M. (2015). A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. *Artificial Intelligence in Medicine, 65*(3), 219–227.

Lai, C., Yeh, W., & Chang, C. (2016). Gene selection using information gain and improved simplified swarm optimization. *Neurocomputing, 218*, 331–338.

Li, M., Han, D., & Wang, W. (2015). Vessel traffic flow forecasting by RSVR with chaotic cloud simulated annealing genetic algorithm and KPCA. *Neurocomputing, 157*, 243–255.

Liang, J., Suganthan, P. N., Chan, C. C., & Huang, V. L. (2006). Wavelength detection in FBG sensor network using tree search DMS-PSO. *IEEE Photonics Technology Letters, 18*, 1305–1307.

Liu, Y., Wang, C., & Zhang, L. (2009). Decision tree based predictive models for breast cancer survivability on imbalanced data. *International conference on bioinformatics and biomedical engineering. IEEE,* 1–4.

Lundin, M., Lundin, J., Burke, H. B., Toikkanen, S., Pylkkänen, L., & Joensuu, H. (1999). Artificial neural networks applied to survival prediction in breast cancer. *Oncology, 57*(4), 281–286.

Martín-Valdivia, M. T., Díaz-Galiano, M. C., Montejo-Raez, A., & Urenã-Lo´pez (2008). Using information gain to improve multi-modal information retrieval systems. *Information Processing & Management, 44*(3), 1146–1158.

Onan, A. (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications, 42*(20), 6844–6852.

Peng, L., Chen, W., Zhou, W., Li, F., Yang, J., & Zhang, J. (2016). An immune-inspired semi-supervised algorithm for breast cancer diagnosis. *Computer Methods and Programs in Biomedicine, 134*, 259–265.

Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support System, 106*, 15–29.

Qiu, H., Yu, H., Wang, L., Yao, Q., Wu, S., Yin, G., et al. (2017). Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy. *Scientific Reports, 7*(1), 16417.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4*(1), 77–90.

Ravdin, P. M., & Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment, 22*(3), 285–293.

Sheikhpour, R., Ghassemi, N., Yaghmaei, P., Ardekani, J., & Shiryazd, M. (2014). Immunohistochemical assessment of p53 protein and its correlation with clinicopathological characteristics in breast cancer patients. *Indian Journal of Science & Technology, 7*(4), 472–479.

Sheikhpour, R., Sarram, M. A., & Sheikhpour, R. (2016). Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. *Applied Soft Computing, 40*, 113–131.

Sizilio, G. R., Leite, C. R., Guerreiro, A. M., & Neto, A. D. D. (2012). Fuzzy method for prediagnosis of breast cancer from the fine needle aspirate analysis. *Biomedical Engineering, 11*(1), 83.

Sørensen, L., & Nielsen, M. (2018). Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination. *Journal of Neuroscience Methods, 302*, 66–74.

Soufan, O., Kleftogiannis, D., Kalnis, P., & Bajic, V. (2015). DWFS: A wrapper feature selection tool based on a parallel genetic algorithm. *PLOS One, 10*(2), E0117988.

Sun, W., Tseng, T. B., Zhang, J., & Qian, W. (2017). Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics, 2017*(57), 4–9.

UCI Machine Learning Repository: Data sets: http://archive.ics.uci.edu/ml/datasets.html.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* NewYork: Springer Verlag.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), 988–999.

Wang, H., Zheng, B., Yoon, S., & Ko, H. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research, 267*(2), 687–699.

Xu, G., Zhou, H., & Chen, J. (2018). CNC internal data based incremental cost-sensitive support vector machine method for tool breakage monitoring in end milling. *Engineering Applications of Artificial Intelligence, 74*, 90–103.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management, 48*(4), 741–754.

Yao, X., & Liu, Y. (1999). Neural networks for breast cancer diagnosis. *Evolutionary computation, 1999. CEC 99. Proceedings of the 1999 congress on. IEEE Xplore, 1767. 3*.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications, 41*(4), 1476–1482.