Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 3875082, 13 pages
https://doi.org/10.1155/2018/3875082

Research Article

Improved Cost-Sensitive Support Vector Machine Classifier for Breast Cancer Diagnosis

Na Liu,[1,2] **Jiang Shen**,[1] **Man Xu**,[3] **Dan Gan**,[1] **Er-Shi Qi**,[1] **and Bo Gao**[4]

[1]*College of Management and Economics, Tianjin University, Tianjin 300072, China*
[2]*School of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China*
[3]*Business School, Nankai University, Tianjin 300071, China*
[4]*School of Computer Science and Technology, Anhui University, Hefei 230601, China*

Correspondence should be addressed to Man Xu; td_xuman@nankai.edu.cn

Received 8 July 2018; Revised 2 October 2018; Accepted 3 October 2018; Published 28 November 2018

Guest Editor: Mustansar A. Ghazanfar

As one of the most prevalent cancers among women worldwide, breast cancer has attracted the most attention by researchers. It has been verified that an accurate and early detection of breast cancer can increase the chances for the patients to take the right treatment plan and survive for a long time. Nowadays, numerous classification methods have been utilized for breast cancer diagnosis. However, most of these classification models have concentrated on maximum the classification accuracy, failed to take into account the unequal misclassification costs for the breast cancer diagnosis. To the best of our knowledge, misclassifying the cancerous patient as non-cancerous has much higher cost compared to misclassifying the non-cancerous as cancerous. Consequently, in order to tackle this deficiency and further improve the classification accuracy of the breast cancer diagnosis, we propose an improved cost-sensitive support vector machine classifier (ICS-SVM) for the diagnosis of breast cancer. In the proposed approach, we take full account of unequal misclassification costs of breast cancer intelligent diagnosis and provide more reasonable results over previous works and conventional classification models. To evaluate the performance of the proposed approach, Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC) breast cancer datasets obtained from the University of California at Irvine (UCI) machine learning repository have been studied. The experimental results demonstrate that the proposed hybrid algorithm outperforms all the existing methods. Promisingly, the proposed method can be regarded as a useful clinical tool for breast cancer diagnosis and could also be applied to other illness diagnosis.

1. Introduction

Breast cancer is one of the most prevalent cancers among women all over the world [1, 2]. According to the American Cancer Society (ACS), an estimation of 252,710 cases have been diagnosed with breast cancer and more than 40,610 women were estimated to die from this cancer in 2017 [3]. In China, breast cancer was listed the sixth leading cause of death among women, and it has been estimated that 214,360 women had died from breast cancer by 2008, and the number of death will reach up to 2.5 million by 2021 [4]. However, it has been verified that an early detection of breast cancer can greatly increase the chances of taking the right decision on a successful treatment plan and ensure a longterm surviving for the patients [5] Consequently, increased attention should be paid to the choice of diagnosis method for the breast cancer. 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 [6]. Therefore, it is absolutely necessary to develop a reasonable scientific method to distinguish malignant lesions from breast tumor lesions.

In recent years, more and more machine learning techniques have been applied for medical diagnosis, which can provide useful knowledge from huge amount of medical data and thereby assist clinical physicians in making correct and effective decisions. To the best of our knowledge, breast cancer diagnosis has been attributed to classification problems. But traditional classification algorithms, such as the Naïve

Bayesian [7], Neural Network [8], Support Vector Machine (SVM) [6], or other hybrid algorithms [9–12], usually aim only to maximize the classification accuracy and fail to take into account the unequal misclassification costs between different categories. However, in most cases, especially in the field of medical diagnosis, the misclassification cost between different categories may vary greatly. Take breast cancer diagnosis as an example; the cost associated with missing a cancer case (false negative) is clearly much higher than that of mislabeling a benign one (false positive). Therefore, standard classifiers inevitably result in an inferior decision making system. In order to overcome this deficiency, in this work we proposed an improved cost-sensitive support vector machine (ICS-SVM) classifier for breast cancer diagnosis, which employs information gain (IG) to select the optimal input feature, which is set to maximize the discrimination capability and fed the selected optimal feature subset into the improved CS-SVM classifier performing for classification. Herein, the motivation for performing feature selection by IG algorithm and SVM performance for classification will be discussed.

To the best of our knowledge, feature selection is the process of selecting the best subset of the input feature to maximize the discrimination capability [13], which seeks to identify significant features and eliminate irrelevant ones to build a good learning model. In this work, the main reason we perform feature selection before breast cancer diagnosis is that it can reduce patients' waiting time without sacrificing the detection accuracy of the breast tumor. Moreover, it could reduce the costs associated with unnecessary biopsies for pathological analyses. Among other feature selection methods, IG has attracted the most attention, which can effectively quantify the correlations between feature and category and it is not sensitive to noise or outlier data [14]. Additionally, it often acts as an important measurement for the features and thereby constructs the optimal feature subset that has the same discriminating ability as the original set of features, leading to a better predictive accuracy. As noted before, in our work, we applied SVM as our underlying classifier, which had two main advantages for breast cancer diagnosis. On the one hand, SVM has a strong generalization performance and classification precision compared with other classification approaches [15]. On the other hand, we can take advantage of its structure and introduce different penalty factors, which we mark C^+ and C^-, respectively, for the positive and negative SVM slack variables during the process of training and testing [16].

In this research, we proposed an improved breast cancer intelligent diagnosis approach, which utilizes IG performance for feature selection and CS-SVM performance for breast cancer classification. We expected that our proposed approach would have a competitive performance for breast cancer diagnosis compared to other classification algorithms. The remainder of this manuscript is organized as follows: Section 2 presents literature review on breast cancer intelligent diagnosis. Section 3 introduces backgrounds and preliminaries of our proposed hybrid algorithm. Section 4 proposes the framework of our proposed approach. Section 5 presents experimental analysis of our proposed algorithm. Section 6 discusses experimental results. Finally, the conclusions of this research are summarized in Section 7.

2. Literature Review

In the literature, various models have been designed to diagnosis breast cancer. In this section, we briefly review the previous studies for breast cancer intelligent diagnosis, such as artificial neural networks (ANNs) and decision tree analysis, which have been utilized for breast cancer diagnosis mostly due to their efficiency and high prediction accuracy. In 2002, Hussein A. Abbass proposed an approach named Memetic Pareto Artificial Neural Network (MPANN) for breast cancer diagnosis, and the experimental results have shown that MPANN has better generalization and lower computational cost than other comparative methods [17]. After that, Marcano-Cedeño, A., J. Quintanilla-Dominguez, and D. Andina proposed an approach called Artificial Metaplasticity Multilayer Perceptron (AMMLP) algorithm to diagnosis of breast cancer, and the experimental results demonstrated that the proposed algorithm can obtain 99.26% classification accuracy, which performed better than other comparative methods [18]. In 2009, Liu et al designed a decision tree prediction model for breast cancer survivability and adopted undersampling method to balance the training data; the results demonstrated that when the ratio is equal to 15%, the AUC of the model is 0.7484 [19]. However, the performance of single learning classification algorithm cannot reflect the interactive factors of the breast cancer survival and recurrence rate [20]. Therefore, in order to overcome the drawbacks brought by single algorithm, many hybrid algorithms have been proposed. In 2009, Akay presented F-score method for feature selection and SVM for breast cancer prediction [21]. On top of that, another hybrid algorithm was presented by Chen et al (2011) that designed a hybrid classifier with rough set for feature selection and SVM for classification [6]. In 2014, Zheng et al. proposed K-means and SVM hybrid algorithm for breast cancer diagnosis, K-means method for breast cancer feature extraction, and SVM for classification [22]. In another study, Onan designed a hybrid intelligent classification model for breast cancer diagnosis, which consists of fuzzy-rough for instance selection, consistency-based for feature selection, and fuzzy-rough nearest neighbor algorithm for breast tumor classification [23]. Additionally, Sheikhpour (2016) proposed PSO and nonparametric kernel density estimation (KDE) based classifier to diagnose breast cancer [24]. To summarize, their results have shown that the proposed hybrid model can achieve high classification accuracy with fewer feature variables. On top of that, some researchers proposed ensemble learning techniques for breast cancer diagnosis. In 2017, Rasti et al. proposed mixture ensemble of convolutional neural networks (ME-CNN) to discriminate between benign and malignant breast tumors, and the experimental results demonstrated that the proposed approach achieved an accuracy of 96.39%, a sensitivity of 97.73%, and a specificity of 94.87%, which has competitive classification performances compared to three existing single-classifier methods and two convolutional ensemble methods [2]. In 2018, Wang et al. designed an ensemble

algorithm fusion SVM for breast cancer diagnosis which emphasizes model structure, and the results demonstrated that the proposed model can achieve the maximum classification accuracy compared to other ensemble models [20]. In summary, as outlined in our literature review, most works focuses on either high classification accuracy or performing feature selection to obtain better data representation for breast cancer diagnosis. However, these studies did not put high emphasis on minimum misclassification cost and maximum classification accuracy with a compact feature subset. Due to this deficiency, in our work, we construct an improved hybrid classifier, which can take full account of feature selection and unequal misclassification costs of breast cancer intelligent diagnosis.

3. Backgrounds and Preliminaries

In this study, we utilized IG algorithm to select a compact feature subset with maximal discriminating ability and applied improved CS-SVM algorithm perform for classification. To the best of our knowledge, when we apply SVM method for classification, whose critical point is how to choose the optimal input feature subset and the optimal parameter, it plays a crucial role in building a classification model with the high classification accuracy and stability [25]. In this work, we applied IG algorithm to select a compact feature subset with the maximal discriminative capability and applied metaheuristic algorithm of simulated annealing particle swarm optimization (SAPSO) to optimize the parameters of CS-SVM classifier. In this regard, this section presents some preliminaries of this hybrid algorithm.

3.1. The Theory of IG Algorithm. In this work, we introduce IG algorithm to select a compact feature subset with the maximal discriminative capability [26]. To the best of our knowledge, the value of IG of each feature can represent its relevance to the category; a higher IG value means that the attribute contributes more information to the category [27]. For the classification system, the information gain of the feature is with respect to class C by counting the number of samples of a feature n in category C [28]. The IG is a measure based on the entropy of a system which can be calculated as follows [29]:

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

There are t classes in the training sample. Let N include n_i samples for class C_i; $i = 1, 2, \cdots, t$; an arbitrary sample belong to class C_i for the possibility of $P(C_i \mid x) = n_i/N$; and N denote the total number of training samples. Thus, the desired information for a given set of training samples is presented as follows:

$$I\left(n_1 + \cdots + n_t\right) = \sum_{i=1}^{t} \frac{n_i}{N} \log_2 \frac{n_i}{N} \tag{2}$$

For any of feature $A = (a_1, a_2, \cdots, a_n)$, the sample set is divided into $(n_1, n_2, \cdots, n_n)$, where n_i contains samples for which sample A has a value of a_i. Let n_i contain n_{ij} samples of class. The expected C_i formation for this division according to attribute A is called the entropy of attribute A.

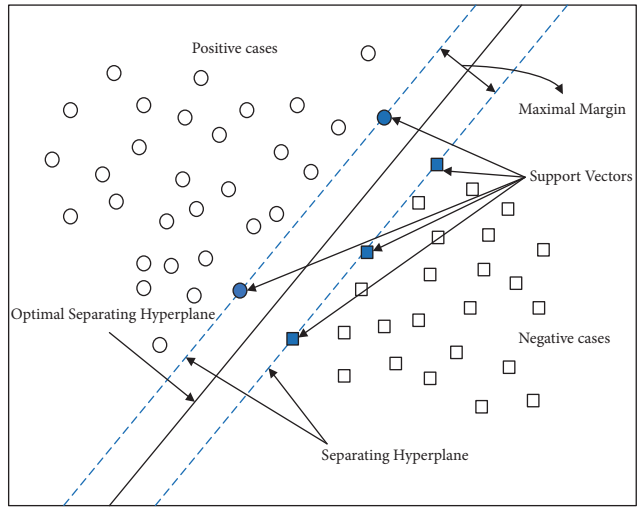

Figure 1: Support vectors and decision boundaries of a linear SVM.

$$E\left(A\right) = \sum_{i=1}^{n} \frac{n_{1i} + \cdots + n_{ti}}{N} I\left(n_{1i} + \cdots + n_{ti}\right) \tag{3}$$

The information gain is defined in (4)

$$InformationGain\left(A\right) = I\left(n_1 + \cdots + n_t\right) - E\left(A\right) \tag{4}$$

3.2. CS-SVM for Classification. SVM was originally developed by Boser and Vapnik, and it has been deemed as an excellent classifier with high generalization ability and structure risk minimum (SRM) for a long time and utilized by many machine learning researchers [30–34]. As can be observed in Figure 1, it presented the support vectors and decision boundaries of SVM classification method. In this paper, we use the LIBSVM toolkit [35] and choose the Radial Basis Function (RBF) kernel.

In our work, we take into consideration the unequal misclassification costs of breast cancer diagnosis and introduce different penalty factors, namely, C^+ and C^- which denote the costs of a false negative and those of a false positive, respectively. Generally, breast cancer diagnosis can be considered as a binary classification problem, where the sample space can be represented as $\{x_i, y_i\}$, $i = 1, 2, \ldots, n$. $y_i \in \{-1, 1\}$. $x_i \in F^d$, x_i are breast tumor samples, and y_i is corresponding label. The samples space has been separated with a hyperplane given by $w^T x + b = 0$, where w is a d-dimensional coefficient vector that is normal to the hyperplane and b is the offset from the original. In our work, we integrated different penalty factors into the

objective function. And the primal problem of CS-SVM has been changed into solving the following optimization task:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C^+ \sum_{i|y_i=+1} \xi_i + C^- \sum_{i|y_i=-1} \xi_i \\ \textit{subject to} \quad & y_i\left[\left(w^T x^{(i)} + b\right)\right] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n \\ & \xi_i \geq 0; \end{aligned} \tag{5}$$

In formula (5), C^+ and C^- denote the unequal misclassification costs of a false negative and those of a false positive, respectively, ξ_i is a slack factor, and b is the threshold for the SVM decision boundary. According to the Lagrange function, the optimization problem can be represented as

$$\begin{aligned} L(w, b, C) = {} & \frac{1}{2}\|w\|^2 + C_i \sum \xi_i \\ & + \sum \alpha\left[1 - \xi_i - y_i\left(w^T x^{(i)} + b\right)\right] \\ & - \sum \beta_i \xi_i \end{aligned} \tag{6}$$

$$C_i = \begin{cases} C & i \in \{i \mid y_i = 1\} \\ C^f \times C & i \in \{i \mid y_i = -1\} \end{cases}$$

In our study, we set benign tumor as negative samples and malignant tumor as positive samples and set the value of false positive much higher than that of false negative and then readjust the parameters of C^+ and C^-, as presented in formula (6). In formula (6), C^f is the corresponding ratio of the two unequal misclassification costs of the breast cancer diagnosis.

In order to solve the problem of formula (6), we set the derivatives of L with respect to w, b, and ξ to be zero and obtain the dual problem as presented in formula (7).

$$\begin{aligned} \arg\max_{\alpha} \quad & L(w, b, \xi) \\ & = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K\left(x^{(i)}, x^{(j)}\right) \\ \textit{subject to} \quad & \begin{cases} \sum_i \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C_i \end{cases} \end{aligned} \tag{7}$$

According to KKT condition, we can convert inequality constraints into equality constraints, and the problem of breast cancer diagnosis based on CS-SVM can be transformed into solving the following minimum objective function:

$$f_o(x) = \operatorname{sign}\left\{\sum_{i=1}^{n} y_i \alpha_i K\left(x * x_i\right) + b_o\right\} \tag{8}$$

In formula (8), $K(x * x_i)$ is a linear kernel function that effectively avoids the dimensionality problem. Considering the breast cancer diagnosis model, $f : X \longrightarrow Y$ is a typical nonlinear modeling problem. Therefore, we use the Radial Basis Function (RBF) kernel function, which exhibits stable performance and other characteristics:

$$K\left(x * x_i\right) = \exp\left(-g\left\|x - x_i\right\|^2\right) \tag{9}$$

In formula (9), $\|x - x_i\|$ is the two-norm distance and g is the kernel function parameter ($g > 0$). Generally, the SVM method using the RBF kernel function must determine two important parameters: the penalty parameter of C and the kernel function parameter of g. To the best of our knowledge, the different parameter pair of (C, g) may have a great influence on the final results. Due to this, in our study, we applied SAPSO algorithm to optimize the parameters of SVM; the details of SAPSO algorithm are presented in Section 3.3.

3.3. SAPSO Algorithm. PSO is a stochastic population-based metaheuristic algorithm which is based on the simulation of the social behavior of organisms, such as birds flying within a flock [36]. Each particle has its own position and velocity to represent its direction and current step, respectively. The position of each particle is modified based on its best position and the best position of the particle swarm [37]. PSO is an efficient optimization algorithm, but the PSO algorithm can easily fall into the local optimum and undergo premature convergence in the global search process. Additionally, the effect of random oscillation is slowed down during the later stage of convergence [38, 39]. Taking this information into account, we present SA algorithm to optimize PSO, which can overcome the drawbacks of PSO algorithm and help the particle jump out of the local optimal and converge to global optimal solution. The formulas of SAPSO algorithm are presented as follows:

$$\begin{aligned} V_{id}(k+1) = {} & \gamma \times \left[V_{id}(k) + c_1 \times r_1\right. \\ & \times \left(P_{id}(k) - X_{id}(k)\right) + c_2 \times r_2 \\ & \left. \times \left(P'_{id}(k) - X_{id}(k)\right)\right] \end{aligned} \tag{10}$$

$$X_{id}(k+1) = X_{id}(k) + V_{id}(k+1) \quad (d = 1, 2, \ldots, n) \tag{11}$$

$$\gamma = \frac{2}{\left|2 - C - \sqrt{C^2 - 4C}\right|}, \quad (C = c_1 + c_2,\ C \geq 4) \tag{12}$$

Formula (10) and formula (11) describe the location and speed update formulas of the particles, and formula (12) describes the convergence factor. $X_{id}(k)$ and $V_{id}(k)$ are the current position and velocity of the particle, respectively; $X_{id}(k+1)$ and $V_{id}(k+1)$ are the updated position and velocity, respectively. c_1 and c_2 are nonnegative constants, called the acceleration factor; r_1 and r_2 are the random numbers belonging to (0, 1); and γ is the convergence factor, which can balance the global and local search capabilities. $P_{id}(k)$ is the local best position, and $P'_{id}(k)$ is the global best position. In SAPSO algorithm, we applied Meteopolis criterion of SA

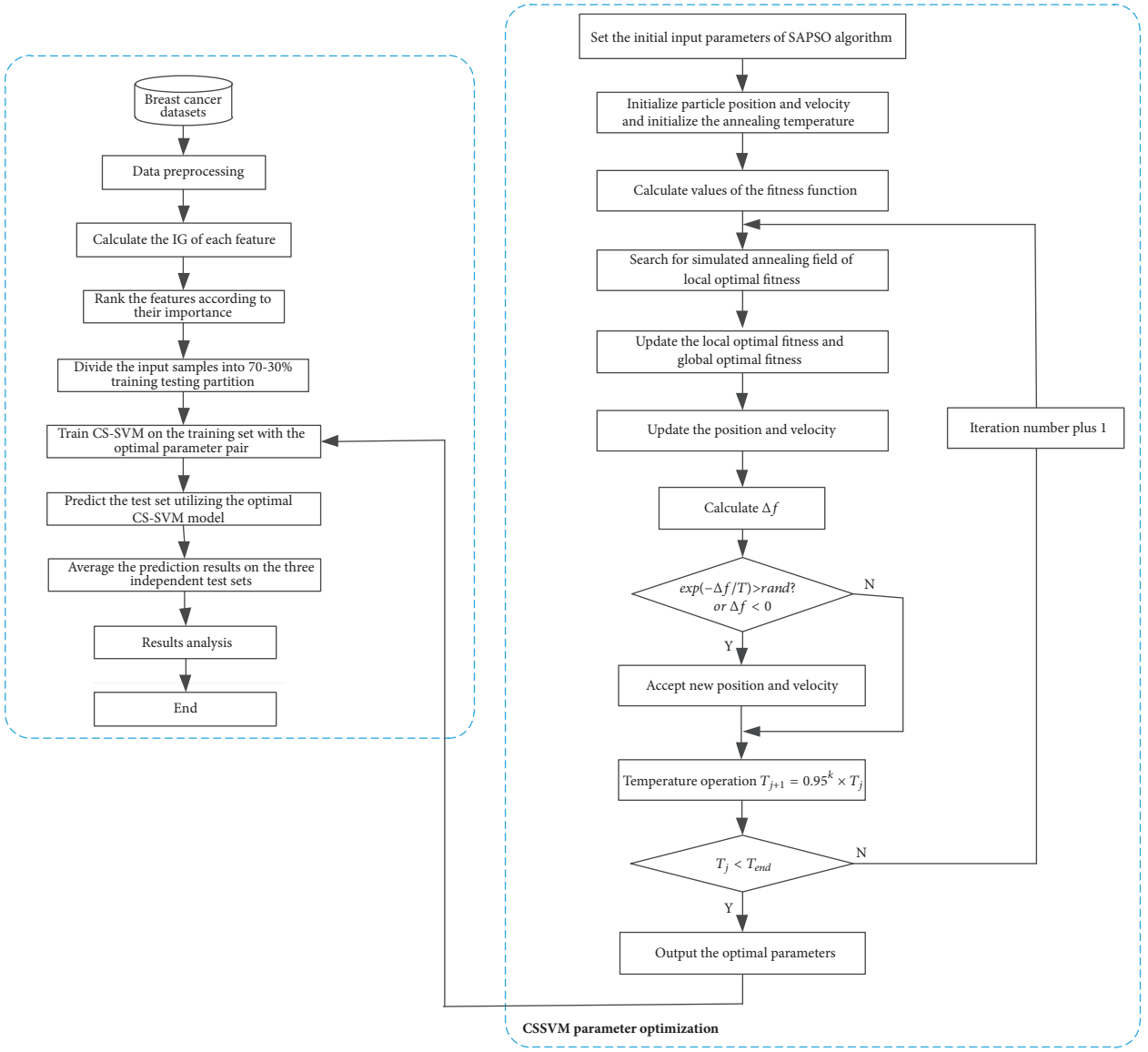

Figure 2: Flowchart of our proposed algorithm.

algorithm to improve PSO and set the acceptance probability of a new solution as follows:

$$P_d = \begin{cases} 1, & f\left(x'\right) < f\left(x\right) \\ \exp\left[-\dfrac{f\left(x'\right) - f\left(x\right)}{T}\right], & f\left(x'\right) \geq f\left(x\right) \end{cases} \tag{13}$$

In formula (13), x is the current solution, and x' is the new solution. $f(\bullet)$ is the fitness function of our algorithm; and T is the current temperature. Additionally, in the iteration process, the annealing temperature follows the criterion below:

$$T = 0.95^k T_0 \tag{14}$$

In formula (14), T is the current temperature and T_0 is the initial temperature. 0.95^k represents the temperature decay coefficient and k is the number of iterations.

4. The Framework of the Proposed Hybrid Algorithm for Breast Cancer Diagnosis

This study proposes a novel breast cancer intelligent diagnosis approach which employs IG algorithm for feature selection and extracting the top n optimal feature utilizing the CS-SVM algorithm, and the resultant of our proposed intelligent diagnosis model can adaptively determine the two key hyperparameters for CS-SVM. The general framework of our proposed method is demonstrated in Figure 2. The proposed model is primarily comprised of three procedures: the feature importance ranking according to IG algorithm, the CS-SVM

Table 1: 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

inner parameter optimization, and the outer classification performance evaluation. At the first step, we rank the features according to their importance and then select the optimized feature subset using CS-SVM classifier. During this process, the parameters of CS-SVM are dynamically adjusted by the SAPSO technique via the 5-fold cross validation analysis. Then, the obtained optimal feature subset and the optimal parameter pair are fed to the prediction model to perform the classification task for breast cancer diagnosis in the outer loop using the 3×5-fold cross validation analysis (i.e., performing 5-fold cross-validation three times).

The classification accuracy and misclassification cost are taken into account in designing the fitness. In this work, we take the average misclassification cost (*AMC*) and average classification error (*ACE)* into consideration and construct the following fitness function:

$$f = avgAMC + avgACE = \frac{\left[\left(\sum_{i=1}^{k} testAMC_i\right) + \left(\sum_{i=1}^{k} testACE_i\right)\right]}{k} \quad (15)$$

$$AMC = \frac{\left(\sum_{i=1}^{k} \left(\left(n_{MB} \cos t_{MB} + n_{BM} \cos t_{BM}\right)/n\right)\right)}{k} \quad (16)$$

$$ACE = \frac{\left(\sum_{i=1}^{k} \left(\left(n_{MB} + n_{BM}\right)/n\right)\right)}{k} \quad (17)$$

where *avgAMC* represents the average misclassification cost achieved by the SVM classifier via 5-fold cross validation and *avgACE* represents the average classification error achieved by the SVM classifier via 5-fold cross validation. n_{MB} represents the number of samples of misclassifying mailignant tumors as benign ones in the test set. $cost_{MB}$ represents the value of misclassification cost of the malignant tumors diagnosed as benign tumors; n_{BM} represents the number of samples of misclassified benign tumors as mailignant ones in the test set. $cost_{BM}$ represents the value of misclassification cost of the benign tumors diagnosed as malignant ones. The cost matrix is presented in Table 1. As for breast cancer diagnosis, the value of $cost_{MB}$ is obviously much higher than that of $cost_{BM}$.

In our study, in order to compare the misclassification costs for the different classification models conveniently, we set the value of the correct classification cost as *0*, and the misclassification cost had to further consider two scenarios: the first scenario is misclassifying malignant tumor as benign ones and missing the best treatment time. The second scenario is misclassifying benign breast tumors as malignant ones resulting in wrong treatment for the patients. The two scenarios as described above may lead to different consequences. The first case may lead the patients to miss the best treatment time and cause the disease to deteriorate, and even worse it may be life-threatening. The second case may lead the patients to take the wrong drugs and cause some side effects. This study completely considers the different scenarios as described above and quantifies the misclassification costs of this two scenarios; then we set $cost_{MB} = 10$ and $cost_{BM} = 1$.

The framework of our proposed hybrid classification algorithm is presented in Figure 2. Herein in this work, we utilized WBC and WDBC breast cancer datasets from UCI machine learning repository. The details of these two datasets are presented in Section 5.1. Here in order to evaluate the performances of our proposed method comprehensively, we set the 70-30% training testing partitions, and the final results are the average results of 3 × 5-fold cross validation with the test set. The main steps of our proposed approach are described below.

Step 1 (data preprocessing). Delete the missing cases and normalize the input samples.

Step 2. Calculate the value of IG for each feature, and rank the features according to their importance.

Step 3. Randomly initialize input samples, and then divide the input samples into 70-30% training testing partitions.

Step 4 (training CS-SVM classifier on the training set with the optimal parameter pair). During this process, we obtained the optimal input parameter pair by using the SAPSO algorithm, and the details of the SAPSO algorithm are presented in Figure 2.

Step 5. Predict the test sets utilizing the optimal CS-SVM model.

Step 6. Average the reported results obtained by 5-fold cross validation.

Step 7. Results' analysis and discussion.

5. Experimental Analysis

In order to examine the effectiveness and rationality of mutual information feature selection method and further to verify the performance of our proposed classification model, we conduct empirical analysis and the experiment was implemented on MATLAB 2016a platform, and the performance parameters of the executing host were Win 10, Inter (R) 1.80 GHz Core (TM) i5-8250U, X64, and 16 GB (RAM).

Table 2: Details of the two datasets.

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

Table 3: Summary of attributes for WBC dataset.

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

5.1. Datasets. To evaluate the performance of the proposed methods, an experiment was conducted based on WBC and WDBC datasets from the UCI repository [40]. The details of the two datasets are presented in Table 2. The class distributions of B and M represent benign tumor and malignant ones, respectively. Additionally, the details of attribute information are presented in Tables 3 and 4.

5.2. Evaluation Measures. To evaluate the performance of our proposed hybrid algorithm, the classification accuracy (ACC), misclassification cost (AMC), and G-mean are utilized as the evaluation approaches. ROC analysis is a widely utilized method for analyzing the performance of binary classifiers and the G-mean is the geometric mean of true positive rate (TPR) and true negative rate (TNR) which is proposed to evaluate the performance of the classifier on imbalanced data. The calculation formulas are presented as follows:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \tag{18}$$

$$AMC = \frac{Number_{FP} \times \cos t_{MB} + Number_{FN} \times \cos t_{BM}}{TP + FN + FP + TN} \tag{19}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{20}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{21}$$

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

The evaluation methods are based on the confusion matrix, which is shown in Table 5. In Table 5, TP is the true positives case, which represents "positive" cases that are correctly classified as "positive." FN represents the "positive" cases that are misclassified as "negative." TN is the true negative cases, which represents "negative" cases that are correctly classified as "negative"; and FP is the false positive cases, which represents "negative" cases that are misclassified as "positive." In this work, we set the positive cases as benign tumors, whereas we set the negative cases as malignant tumors.

5.3. Experimental Procedure. This section presents the detailed experimental procedure of our proposed approach. In the experiments, we utilized 3 × 5-fold cross validation method to obtain the final results. In each fold, training dataset is first fed to the IG feature selection algorithm, generating different feature subsets. In this work the results of IG for the two datasets are presented in Tables 6 and 7. Training dataset with the selected feature subset is then used to train outer CS-SVM classifier, and during this process the best parameter pairs of CS-SVM algorithm were obtained by the inner SAPSO algorithm. The main parameters of SAPSO algorithm are presented in Table 8. Finally, the performance of the test dataset is obtained for evaluating our selected features by the optimal CS-SVM model. The final results have been achieved by averaging the results of 5-fold cross validation method with the test set. In this scheme, the test datasets are randomly partitioned into 5 equal sized partitions. Each time, one of the partitions is used for validation and the remainder of the partitions is utilized for training. This process is repeated five times and the average results are reported. As can be observed from Figures 3 and 4, it presents the best results of our proposed method based on WBC and WDBC datasets. And Figures 5 and 6 present the ROC of our proposed approach. From these two figures, we can obviously see that the results of our proposed approach can achieve the promising results for breast cancer diagnosis.

In order to verify the superiority of our proposed method, we conduct two test sequences. First, our approach was compared with some previous works proposed by other authors; the results are presented in Table 9. The second sequence is to compare our proposed method with some conventional classification approaches. The results of these classification models are presented in Table 10. The source codes employed in this work for all the conventional classification models were implemented in MATLAB 2016a, with the help of LIBSVM toolkit [35]. In addition, to examine the performances of different classification models, ACC, AMC, G-mean, sensitivity, and specificity are utilized as the evaluation measures. Moreover, in order to verify the superiority of feature selection proposed in our approach, we also compare the running time of our proposed method with those of other two classification models and the results are presented in Figure 7.

Table 4: Summary of attributes for WDBC dataset.

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

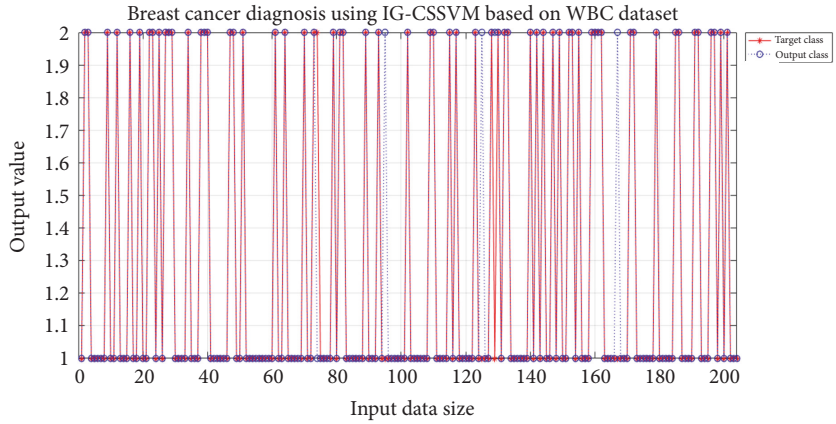

Figure 3: Classification results of our proposed method with WBC test set.

Table 5: Confusion matrix.

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

Table 6: The order of features based on IG for WBC dataset.

Rank *No.*	Feature name
1	Mitoses
2	Clump Thickness
3	Marginal Adhesion
4	Normal Nucleoli
5	Single Epithelial Cell Size
6	Bland Chromatin
7	Bare Nuclei
8	Uniformity of Cell Shape
9	Uniformity of Cell Size

5.4. Experimental Results and Analysis. The final results are obtained by 3 × 5-fold cross validation method, and the results are reported by the average results of the 5-fold cross validation. As can be observed form the results listed, the proposed classification model is the most suitable method for breast cancer diagnosis, which obtains promising results and achieves classification accuracies of 98.04% for the WBC dataset and 98.83% for the WDBC dataset. In order to assess the actual situation of breast cancer diagnosis, we introduce misclassification cost as the measurement to evaluate the unequal costs of different categories for breast cancer diagnosis. As can be observed from the results listed in Table 9, our proposed approach performs well for breast cancer diagnosis, which takes full account of the unequal misclassification costs and obtains promising results for breast cancer diagnosis. As can be observed in Table 9, it obviously indicates that our proposed approach obtains the promising results, which make the breast cancer intelligent diagnosis more reasonable than previous literatures. As can be observed in Table 10, it presents the classification results of different conventional classification models, which obviously indicate that our proposed method has superior performances compared to other conventional classification methods, which obtains the classification accuracy of 98.04% for the WBC dataset and 98.83% for the WDBC dataset. The details of the results are shown in Figures 3 and 4, where, in these two figures, symbols of "1" and "2" represent benign tumors and malignant ones, respectively. To further evaluate the effectiveness of the proposed approach, we applied the performance of ROC curve for evaluation, and the results of area under the curve (AUC) are shown in Figures 5 and 6. Form these two figures we can see that the AUC of WBC and

Table 7: The order of features based on IG for WDBC dataset.

Rank *No.*	Feature name	Rank *No.*	Feature name	Rank *No.*	Feature name
1	Texture 04	*11*	symmetry 01	*21*	smoothness 01
2	Radius 04	*12*	area 01	*22*	perimeter
3	Compactness 06	*13*	concave points	*23*	area 03
4	Concavity 01	*14*	compactness 05	*24*	fractal dimension 01
5	radius 02	*15*	symmetry 03	*25*	perimeter
6	smoothness 02	*16*	compactness 04	*26*	compactness 03
7	symmetry 02	*17*	concavity 02	*27*	concavity 03
8	fractal dimension 02	*18*	texture 03	*28*	compactness 01
9	radius 01	*19*	radius 03	*29*	texture 02
10	texture 01	*20*	compactness 02	*30*	area 02

Table 8: The main parameters of SAPSO algorithm.

Parameter	Value
Objective function	Fitness value obtained by formula (15)
Maximum number of evolutions	300
maximum number of populations	20
Acceleration factor of c_1	1.9
Acceleration factor of c_2	1.9
The maximum value of individual speed	0.5
The minimum value of individual speed	-0.5
Initial temperature	100

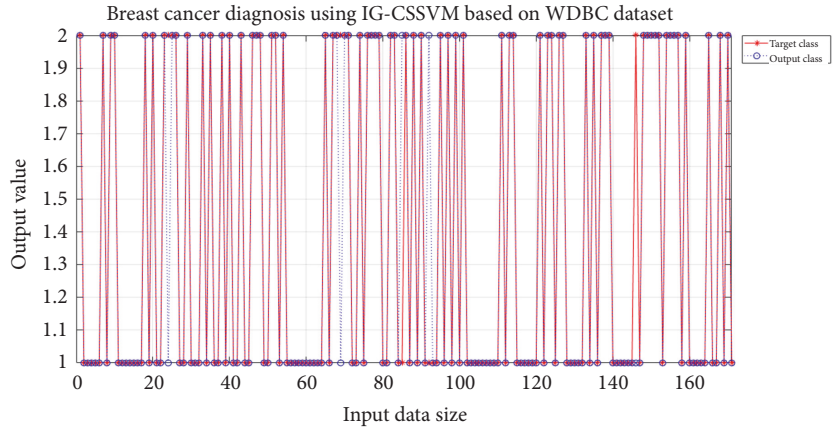

Figure 4: Classification results of our proposed method with WDBC test set.

WDBC datasets are 0.9564 and 0.9578, respectively, which yield promising results on these two breast cancer datasets. Moreover, to confirm the superiority of feature selection of our proposed approach, we also compare the running time of our proposed approach with other two comparison methods and the results are presented in Figure 7. From the experimental results we can obviously deduce that our proposed approach has higher calculation efficiency than other two comparative models.

6. Discussion

In this section, we will provide a discussion on the performance of our proposed approach. The proposed approach is a hybrid method based on the IG for feature selection and the ICS-SVM performance for classification. As mentioned in advance, algorithm to be utilized in the feature selection stage, the classifier to be utilized for classification, and the meta-heuristic algorithm to be employed for searching the optimal parameter pairs of the proposed classifier are essential factors in building our classifier.

In this regard, an extensive experimental analysis has been implemented on WBC and WDBC breast cancer datasets. In order to evaluate the performances of our proposed approach, we implemented two test sequences. The first is to compare our proposed method with some previous works, and the results are presented in Table 9, from Table 9

Table 9: Comparison of our proposed approach with previous works.

Author	Year	Model	Dataset	ACC(%)	AMC	G-mean(%)	Sen(%)	Spec(%)
Karabatak[41]	2009	Neural network classification with association rules for reducing the dimension.	WBC	95.60	—	—	—	—
Zheng[22]	2014	Support vector machine algorithms with K-means for feature extraction	WBC	97.38	—	—	—	—
Nahato[42]	2015	Rough set indiscernibility relation method and the backpropagation neural network	WBC	98.61	—	98.60∗	98.76	98.57
Wang[20]	2018	SVM-based ensemble learning algorithm	WBC	97.10	—	97.17	97.11	97.23
			WDBC	97.68	—	97.09	94.75	99.49
Proposed	—	Cost-sensitive SVM with IG for feature selection	WBC	98.74∗	0.064	98.13	97.88	98.38
			WDBC	98.83∗	0.129	97.35	99.01∗	95.71

Note: the symbol of "∗" represent the optimal value for each performance.

Table 10: Comparison of our proposed method with conventional classification models.

Method	Dataset	ACC(%)	AMC	G-mean(%)	Sen(%)	Spec(%)
SVM(RBF)	WBC	96.58	0.132	96.67	96.35	97.05
	WDBC	95.91	0.251	95.15	97.37	92.98
PSO-SVM(RBF)	WBC	95.61	0.307	94.37	97.82	91.04
	WDBC	97.66	0.234	97.05	100∗	94.20
BP neural network	WBC	94.11	0.324	92.20	93.30	91.30
	WDBC	94.72	0.526	92.93	100∗	86.41
LVQ neural network	WBC	91.56	0.627	87.88	96.55	80.00
	WDBC	92.75	0.724	90.26	100∗	81.48
3-NN	WBC	91.10	0.485	90.90	92.50	89.30
	WDBC	92.60	0.602	91.42	97.50	85.71
Decision Tree	WBC	96.64	0.162	96.63	97.84	95.45
	WDBC	95.65	0.304	95.27	97.50	93.10
Random Forest	WBC	96.49	0.140	96.49	96.33	96.66
	WDBC	97.53	0.145	96.82	97.82	95.83
Proposed	WBC	98.04∗	0.064∗	98.13∗	97.88	98.38∗
	WDBC	98.83∗	0.129∗	97.35∗	99.01	95.71

Note: the symbol of "∗" represent the optimal results for each performance.

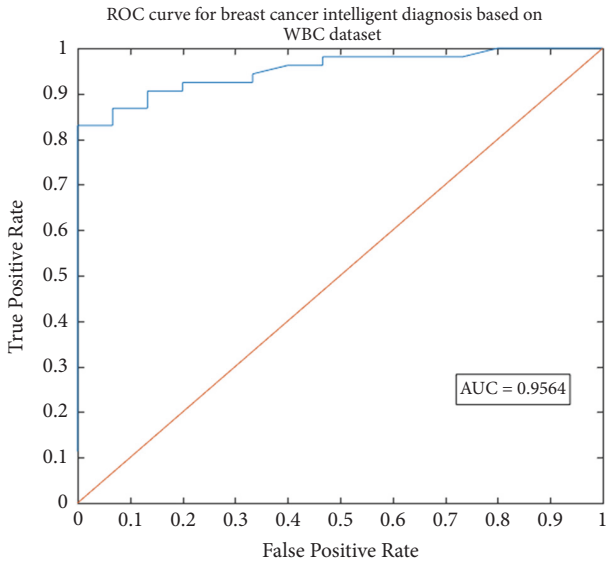

Figure 5: ROC curve of our proposed intelligent classification method based on WBC dataset.

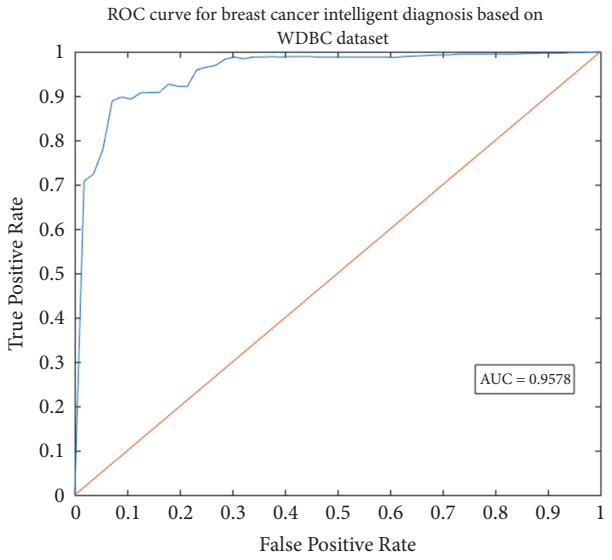

Figure 6: ROC curve of our proposed intelligent classification method based on WDBC dataset.

we can obviously see that our proposed approach can achieve best performances in terms of ACC and AMC. Additionally, from Table 9 we can obviously see that the previous works have not considered the unequal misclassification cost for breast cancer diagnosis, and it is not inconsistent with the actual situation for breast cancer diagnosis. Consequently, in our proposed approach we take full account of misclassification cost and construct an improved CS-SVM classifier for breast cancer diagnosis, and the results from Table 9 verified that our proposed approach can be utilized as the best classifier for breast cancer intelligent diagnosis. In order to verify the superior performances of our proposed approach, we also compared our proposed method with some conventional classification methods, and the results are presented in Table 10. Form Table 10, we can see that when we compared our proposed approach with SVM(RBF), PSO-SVM(RBF), BP neural network, LVQ neural network, 3-NN, decision tree, and random forest methods, the results demonstrated that our proposed approach can obtain the optimal values in terms of ACC, AMC, and G-mean. In the experimental analysis, we implement 3×5-fold cross validation method, and the best results of classification accuracy are 98.04% for WBC dataset and 98.83% for WDBC dataset. And the results of the highest classification accuracies are presented

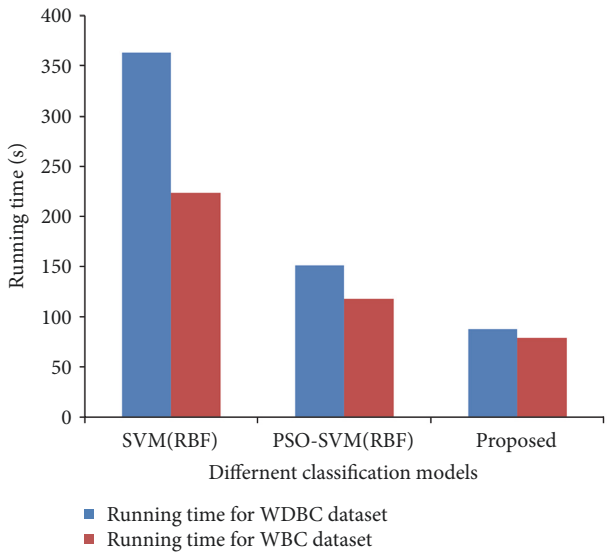

Figure 7: Comparison of running time for different classification models.

in Figures 3 and 4, and the corresponding ROC curves are presented in Figures 5 and 6; from these two figures, we can obviously see that the AUC of our proposed approach based on WBC and WDBC datasets are 0.9564 and 0.9578, respectively, which achieved promising results for breast cancer diagnosis.

To highlight the importance of utilizing IG for feature selection in our proposed approach, we also compared our proposed method with other two alternative methods. One is utilizing the grid-search method to handle the SVM's parameter pair, and the other is utilizing the PSO algorithm to optimize the SVM's parameter pair. The results of these comparison methods are presented in Table 10, and the running time of these three models is presented in Figure 7. From the results of these three models, we can deduce that utilizing IG for feature selection can decrease the dimension of feature space and improve the computation efficiency.

From the empirical results based on WBC and WDBC breast cancer datasets, we can deduce that our proposed approach is the most suitable method for breast cancer diagnosis, which can produce excellent performances and only requires a minimum computational cost for solving breast cancer classification problem. Promisingly, our proposed approach may be adapted to other diseases' diagnosis.

7. Conclusion

In this study, an improved CS-SVM classifier is proposed for breast cancer diagnosis. The proposed approach not only takes full account of unequal misclassification cost of breast cancer diagnosis, but also employs IG for features selection and utilizes meta-heuristic method to optimize the classifier. In order to verify the performances of our proposed approach, the proposed improved classifier was evaluated by several experiments on WBC and WDBC datasets and the experimental results demonstrated that our proposed approach can yield promising results for breast cancer diagnosis in comparison to some previous works and conventional classification methods (i.e., SVM (RBF), PSO-SVM (RBF), BP neural network, LVQ neural network, 3-NN, decision tree, and random forest). The main objective of this work was to construct an effective classifier for breast cancer diagnosis and expect our research to be utilized in real clinical diagnostic system and thereby assist clinical physicians in making correct and effective decisions in the future.

Data Availability

All the datasets we utilized in this paper are all come from the UCI Machine Learning Repository.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (71571105).

References

[1] R. Sheikhpour, N. Ghassemi, P. Yaghmaei, J. M. Ardekani, and M. Shiryazd, "Immunohistochemical assessment of p53 protein and its correlation with clinicopathological characteristics in breast cancer patients," *Indian Journal of Science and Technology*, vol. 7, no. 4, pp. 472–479, 2014.

[2] R. Rasti, M. Teshnehlab, and S. L. Phung, "Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks," *Pattern Recognition*, vol. 72, pp. 381–390, 2017.

[3] American Cancer Society (ACS), *Breast cancer facts & figures*, American Cancer Society, 2018.

[4] L. Fan, K. Strasser-Weippl, J. J. Li et al., "Breast cancer in China," *The Lancet Oncology*, vol. 15, pp. 279–289, 2014.

[5] G. R. M. A. Sizilio, C. R. M. Leite, A. M. G. Guerreiro, and A. D. D. Neto, "Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis," *Biomedical Engineering Online*, vol. 11, article no. 83, 2012.

[6] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 38, no. 7, pp. 9014–9022, 2011.

[7] B. Krawczyk, G. Schaefer, and M. Woźniak, "A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification," *Artificial Intelligence in Medicine*, vol. 65, no. 3, pp. 219–227, 2015.

[8] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using Genetically Optimized Neural Network model," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4611–4620, 2015.

[9] H. Ahn and K.-J. Kim, "Global optimization of case-based reasoning for breast cytology diagnosis," *Expert Systems with Applications*, vol. 36, no. 1, pp. 724–734, 2009.

[10] L. Peng, W. Chen, W. Zhou, F. Li, J. Yang, and J. Zhang, "An immune-inspired semi-supervised algorithm for breast cancer

diagnosis," *Computer Methods and Programs in Biomedicine*, vol. 134, pp. 259–265, 2016.

[11] W. Sun, T.-L. B. Tseng, J. Zhang, and W. Qian, "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data," *Computerized Medical Imaging and Graphics*, vol. 57, pp. 4–9, 2017.

[12] D. Gu, C. Liang, and H. Zhao, "A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 77, pp. 31–47, 2017.

[13] Z. Wang, M. Li, and J. Li, "A multi-objective evolutionary algorithm for feature selection based on mutual information with a new redundancy measure," *Information Sciences*, vol. 307, pp. 73–88, 2015.

[14] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825–1844, 2007.

[15] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proceedings of the 15th European Conference on Machine Learning (ECML '04)*, vol. 3201 of *Lecture Notes in Computer Science*, pp. 39–50, September 2004.

[16] G. Xu, H. Zhou, and J. Chen, "CNC internal data based incremental cost-sensitive support vector machine method for tool breakage monitoring in end milling," *Engineering Applications of Artificial Intelligence*, vol. 74, pp. 90–103, 2018.

[17] H. A. Abbass, "An evolutionary artificial neural networks approach for breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 25, no. 3, pp. 265–281, 2002.

[18] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9573–9579, 2011.

[19] Y. Liu, C. Wang, and L. Zhang, "Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data," in *Proceedings of the 2009 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*, pp. 1–4, Beijing, China, June 2009.

[20] H. Wang, B. Zheng, S. W. Yoon, and H. . Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *European Journal of Operational Research*, vol. 267, no. 2, pp. 687–699, 2018.

[21] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240–3247, 2009.

[22] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.

[23] A. Onan, "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*, vol. 42, no. 20, pp. 6844–6852, 2015.

[24] R. Sheikhpour, M. A. Sarram, and R. Sheikhpour, "Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer," *Applied Soft Computing*, vol. 40, pp. 113–131, 2016.

[25] H. Frohlich, O. Chapelle, and B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm," in *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 142–148, Sacramento, Calif, USA, 2003.

[26] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996.

[27] J. M. Yang, Y. N. Liu, X. D. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Information Processing & Management*, vol. 48, no. 4, pp. 741–754, 2012.

[28] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.

[29] D. Huang and T. W. S. Chow, "Effective feature selection scheme using mutual information," *Neurocomputing*, vol. 63, pp. 325–343, 2005.

[30] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "Training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory (COLT '92)*, pp. 144–152, July 1992.

[31] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 10, no. 5, pp. 988–999, 1999.

[32] H. Yuan, J. Chen, and G. Dong, "Bearing Fault Diagnosis Based on Improved Locality-Constrained Linear Coding and Adaptive PSO-Optimized SVM," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[33] D. Cui and K. Xia, "Strip Surface Defects Recognition Based on PSO-RS&SOCP-SVM Algorithm," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[34] F. Cheng, Y. Zhou, J. Gao, and S. Zheng, "Efficient Optimization of F-Measure with Cost-Sensitive SVM," *Mathematical Problems in Engineering*, vol. 2016, Article ID 5873769, pp. 1–11, 2016.

[35] C. Chang and C. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[36] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, Perth, Australia, December 1995.

[37] E. Talbi, "Metaheuristics: From Design to Implementation. Proceedings of SPIE," *The International Society for Optical Engineering*, vol. 42, no. 4, pp. 497–541, 2009.

[38] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[39] M. Tao, S. Huang, Y. Li, M. Yan, and Y. Zhou, "SA-PSO based optimizing reader deployment in large-scale RFID Systems," *Journal of Network and Computer Applications*, vol. 52, pp. 90–100, 2015.

[40] D. Dheeru and E. Karra Taniskidou, *UCI machine learning repository*, http://archive.ics.uci.edu/ml.

[41] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3465–3469, 2009.

[42] K. B. Nahato, K. N. Harichandran, and K. Arputharaj, "Knowledge mining from clinical datasets using rough sets and back-propagation neural network," *Computational and Mathematical Methods in Medicine*, vol. 2015, 2015.