European Journal of Operational Research 000 (2018) 1–13

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

A support vector machine-based ensemble algorithm for breast cancer diagnosis

Haifeng Wang[a], Bichen Zheng[a], Sang Won Yoon[a,*], Hoo Sang Ko[b]

[a] Department of Systems Science and Industrial Engineering, State University of New York at Binghamton, Binghamton, NY 13902, United States
[b] Department of Mechanical and Industrial Engineering, Southern Illinois University Edwardsville, Edwardsville, IL 62026, United States

ARTICLE INFO

Article history:
Received 19 July 2016
Accepted 1 December 2017
Available online xxx

Keywords:
Analytics
Cancer diagnoses
Support vector machine
Ensemble learning
Variance reduction

ABSTRACT

This research studies a support vector machine (SVM)-based ensemble learning algorithm for breast cancer diagnosis. Illness diagnosis plays a critical role in designating treatment strategies, which are highly related to patient safety. Nowadays, numerous classification models in data mining domains are adapted to breast cancer diagnosis based on patients' historical medical records. However, the performance of each algorithm depends on various model configurations, such as input feature types and model parameters. To tackle the limitation of individual model performance, this research focuses on breast cancer diagnosis that uses an SVM-based ensemble learning algorithm to reduce the diagnosis variance and increase diagnosis accuracy. Twelve different SVMs, based on the proposed Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) approach, are hybridized. To evaluate the performance of the proposed model, Wisconsin Breast Cancer, Wisconsin Diagnostic Breast Cancer, and the U.S. National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program breast cancer datasets have been studied. The experimental results show that the WAUCE model achieves a higher accuracy with a significantly lower variance for breast cancer diagnosis compared to five other ensemble mechanisms and two common ensemble models, i.e., adaptive boosting and bagging classification tree. The proposed WAUCE model reduces the variance by 97.89% and increases accuracy by 33.34%, compared to the best single SVM model on the SEER dataset. In practice, the proposed methodology can be further applied to other illness diagnoses, which offers an alternative to a safer, more reliable, and more robust illness diagnosis process.

1. Introduction

In 2011, a survey of about 187 countries' breast cancer mortality and incidence rates from 1980 to 2010 indicated that global breast cancer incidences increased from 641,000 cases in 1980 to 1,643,000 cases in 2010, with an average annual increase rate of 3.1% (Forouzanfar et al., 2011). In particular, breast cancer is one of the leading cancer instances for women, which contributes 15% to total cancer deaths in 2015 in the United States (Siegel, Miller, & Jemal, 2015).

To tackle the dramatically increasing cancer rate, early detection approaches are widely discussed in many disease prevention studies. Commonly used pre-diagnosis methods include annual mammography (Ades et al., 2014), gene diagnosis (Liu & Sotiriou, 2002), clinical diagnosis (Sotiriou et al., 2003), etc. Moreover, with the development of biomedical technologies and information technologies in recent years, various prognostic factors related to breast cancer have been recorded, which enabled many researchers to develop more sophisticated early detection models using different data driven prediction methodologies, such as support vector machines (SVMs), logistic regression (LR), multilayer perceptrons (MLPs), and decision trees (DTs).

In the domain of data mining, breast cancer prediction problems are also considered as classification problems to classify benign and malignant tumors, which use various breast tumor measurements instead of conventional diagnostic lab tests, such as breast biopsy, positron emission tomography, and magnetic resonance imaging (Gupta, Kumar, & Sharma, 2011; Zheng, Yoon, & Lam, 2014). The classification performance depends on many factors, such as input features, parameter settings, and model structures. It is still challenging to find an effective strategy to obtain a good performance for general classification problems (Friedrichs & Igel, 2005; Zheng et al., 2015). In particular, breast cancer diagnosis is more important than ever because the classification results directly affect patients' treatment and safety. It requires not only a high prediction of accuracy, but also a high reliability and robustness, which is another challenge for data mining researchers.

* Corresponding author.
E-mail address: yoons@binghamton.edu (S.W. Yoon).

https://doi.org/10.1016/j.ejor.2017.12.001
0377-2217/

It is obvious that each algorithm has its advantages and limitations over different classification tasks (Wolpert, 2002). One of the most popular methods is ensemble learning to leverage the strength of individual classifiers. While a good ensemble classifier can be built upon a number of weak bases, studies show that the property of base classifiers impacts the effectiveness of the ensemble outputs (Breiman, 1996). In this research, to compensate for the limitations and maximize the advantages of individual base classifiers, multiple structures of SVM models, known for high classification accuracy, are adopted to be hybridized as an ensemble learning model for breast cancer diagnosis. The proposed ensemble model includes two types of SVM structures, i.e., a *C*-SVM and a ν-SVM, and six types of kernel functions. To import the expertise of different base classifiers on diagnostic tasks, a Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) mechanism is proposed for model hybridization. The proposed model is further validated and evaluated based on two standard breast cancer datasets, the Wisconsin Breast Cancer (WBC) dataset and the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, and one practical large scale dataset, the Surveillance, Epidemiology, and End Results (SEER) dataset, which was released by the U.S. National Cancer Institute through a cancer statistics program (SEER, 2017).

The rest of the paper is organized as follows: Breast cancer diagnosis studies, related SVM, and ensemble learning methodologies are reviewed and summarized in Section 2. Section 3 provides a detailed discussion on the proposed method in terms of ensemble structure and diagnosis accuracy. By investigating SVM accuracy surfaces, it is confirmed that the proposed model paradigm is quite necessary to overcome the influence of single model parameter settings. Experimental results and analysis are presented in Section 4. Conclusions and possible future research directions are explored in Section 5.

2. Literature review

In the literature, various models have been designated to identify breast cancer cases based on recorded clinical features, such as tumor sizes, texture behaviors, and uniformity of cell shapes. Among these methodologies, data-oriented machine learning models provide a low cost clinical examination assistance for breast cancer diagnosis compared to other lab tests. In this section, common data-oriented classifiers in the area of breast cancer diagnosis are first summarized. SVM, as one of the effective classification method with high generalization performance as shown in many studies, is reviewed in the following subsection. Finally, recent ensemble learning techniques are further introduced for improving the classification accuracy over individual models.

2.1. Common classifiers for breast cancer diagnosis

Many basic data mining models, such as artificial neural networks (ANNs), and DT models, have been applied to breast cancer diagnosis due to their cost effectiveness and high accuracy. Ravdin and Clark (1992) utilized neural network (NN) models to predict patient survival chances at different future time points by involving time factors in prognostic variables. The performance of the NN was compared to a regression model on 1373 patients' censored survival data, and achieved a similar accuracy level. Mangasarian, Street, and Wolberg (1995) designed a linear programming-based diagnostic system to predict malignant probabilities for nonrecurring cases and recurring time for recurring cases. Their model was tested on 569 patients using a cross-validation approach, and obtained 97.5% predicted accuracy. On top of a modified C4.5 decision tree algorithm, Quinlan (1996) further improved the C4.5 classifier accuracy to 94.74% by incorporating the Minimum Description Length (MDL) penalty. By comparing a DT model (i.e., C5) with ANNs and LR on a large dataset (more than 200,000 cases), Delen, Walker, and Kadam (2005) found that the C5 model can achieve the best prediction accuracy by 93.6% more than the others for the holdout large dataset.

Performance of a single learning algorithm can be impacted by different factors, such as feature space characteristics, algorithm parameters, and solution approaches. To systematically configure these parameters, many hybrid models have been proposed. Ravi and Zimmermann (2000) combined feature selection and fuzzy systems together to select the critical features from a dataset for model performance improvement. In their model, a modified threshold accepting algorithm was proposed to minimize the number of rules during model training process. Their model was tested on wine classification and the Wisconsin breast cancer determination problems. Their results show that the proposed model can achieve high classification powers when working with fewer feature variables. In addition to the hybrid with feature selection methods, feature extraction methods, such as principle component analysis (PCA), were also applied in their later work (Ravi, Reddy, & Zimmermann, 2000). To improve the prediction performance for a large dataset, a fuzzy decision tree (FDT), a hybrid DT and fuzzy rules, was proposed to estimate breast cancer recurrence for patients (Khan, Choi, Shin, & Kim, 2008). In their proposed model, a DT approach was used to generate fuzzy rules for a fuzzy inference system. The model was tested on the SEER dataset. It was noticed that the FDT model was more robust than individual models (Khan et al., 2008). By hybridizing NN models and association rules, Karabatak and Ince (2009) proposed an automatic breast cancer diagnostic system. In their model, association rules were utilized to reduce feature space and NN was used for classification. The model was tested on the WBC dataset, and their results show that the hybrid diagnostic system outperformed the single NN model on both effectiveness and efficiency (Karabatak & Ince, 2009).

In most of the cases, model parameter setting is one of the critical challenges that impacts model performance. Some works have also been conducted to combine meta-heuristic algorithms with data mining models to help model parameter tuning. To enhance the wavelet neural network training process, Chauhan, Ravi, and Chandra (2009) proposed a differential evolution trained wavelet neural network (DEWNN), where a differential evolution method was used to search the best parameter settings for a wavelet neural network. The DEWNN model was tested on three bank bankruptcy datasets and three standard datasets that included the WBC dataset. Their results show that DEWNN can achieve relatively high generalization ability. Naveen, Ravi, Rao, and Chauhan (2010) also combined the differential evolution method with other two algorithms, i.e., *K*-means, which performs data point centralization, and a radial basis function (RBF) network, which is used as supervised learning. Their proposed model, differential evolution trained radial basis function (DERBF) network, was compared to several existing models, such as DEWNN and threshold accepting trained wavelet neural network, on bank bankruptcy datasets and standard datasets, and obtained fairly high accuracy.

In summary, most studies in the literature focused on either tuning model parameters, or performing feature extraction and/or selection to obtain better data representations. However, these studies did not put high emphasis on model structures, which might significantly affect the performance. Moreover, the parameter tuning process, such as using differential evolution, may also lead to a potential risk of overfitting.

2.2. Support vector machine for breast cancer diagnosis

To account for model structure characteristics and overfitting issues, several novel model structures have been developed over the

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001

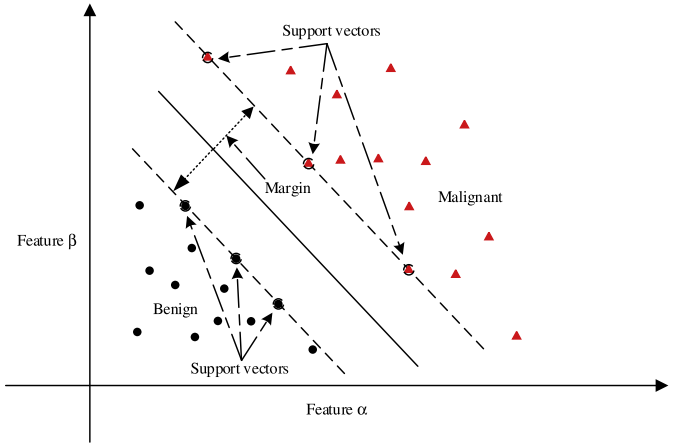


Fig. 1. Structure of SVM on a two dimensional feature space.

last few decades. SVM, a group of margin classifier models proposed by Vapnik and his research group at AT&T Bell Laboratories in the 1990s, is one type of the effective models with high generalization ability in practice (Cortes & Vapnik, 1995). Different from empirical risk minimization-based statistical learning methods, SVM aims to minimize structural risk, which demonstrates a strong capability in overfitting avoidance (Ayat, Cheriet, & Suen, 2005). In the SVM model, decision hyperplanes are constructed based on identified support vectors to form a separation gap to divide two class instances with the maximal margin, as shown in Fig. 1.

Due to a widespread generalization ability of SVMs compared to conventional learning approaches, SVMs have been applied to many fields. In particular, as a data driven prediction technique, SVM models have attracted the most attention on illness diagnosis in recent years, such as cerebral palsy gait diagnosis, gastric lymph node cancer detection, and prostate cancer diagnosis (Ishikawa, Takahashi, Takemura, Mizoguchi, & Kuwata, 2014; Kamruzzaman & Begg, 2006; Shah et al., 2012; Son, Kim, Kim, Choi, & Lee, 2010). SVM has been utilized for breast cancer diagnosis since the 1990s. Using the WBC dataset, Liu et al. (2003) analyzed the impact of SVM kernel functions and parameters on classification accuracy. Compared to *K*-means clustering, MLP, and probability neural network (PNN), Liu et al. (2003) concluded that SVM had the best performance in diagnosing clinic breast cancer. Another research work on diagnosis of breast cancer recurrence also demonstrated the outstanding performance of SVM over ANN, Cox-proportional hazard regression model, and three other prognostic models (Kim et al., 2012), where the validation experiments were performed on 679 patient cases, those who underwent breast cancer surgery between 1994 and 2002 in Korea.

Although the SVM technique has shown its superiority in accurately diagnosing breast cancer, the performance can be further improved by optimizing its structures and parameter configurations. A Mixture of Rough set and Support vector machine (MRS) model, proposed by Zeng and Liu (2010), reorganized sample space by rough set (RS) theory. Their model included two layers: the first layer applied RS theory to identify singular samples, and the second layer utilized an SVM model to classify the remaining samples. The MRS model was extended for the case by applying RS theory to remove redundant features in the feature selection process by Chen, Yang, Liu, and Liu (2011). As a result, the number of features was reduced from ten to five on the WBC dataset. Instead of using feature selection, Zheng et al. (2014) proposed a *K*-SVM model by hybridizing a *K*-means algorithm with SVM on breast cancer datasets (Zheng et al., 2014). In their model, *K*-means was utilized to extract the abstract malignant and benign patterns. A similarity index, based on the comparison between the samples and the abstract patterns, was derived. The feature space for the SVM learning was reconstructed based on these similarity indexes. Their model achieved the highest accuracy, i.e., 97.38%, with the least CPU time on the WDBC dataset as compared to other heuristic-based feature selection methods (Zheng et al., 2014).

Although SVM obtains overall good performance, there are still some nontrivial parameters, such as kernel types, regularization parameters, and kernel parameters, which affect the performance of the individual SVM model. Heuristic techniques, such as gradient-based techniques (Chapelle, Vapnik, Bousquet, & Mukherjee, 2002), evolutionary algorithms (Lorena & De Carvalho, 2008), and particle swarm optimization (Gomes, Prudêncio, Soares, Rossi, & Carvalho, 2010), have been put into the SVM parameter selection procedure. Even though the heuristic optimization process fine tunes the parameter values for SVM models, the search process is time consuming and may terminate with local optima. It also cannot exceed the structure limitation of an individual SVM.

2.3. Ensemble learning for classification

Ensemble learning was initially proposed for reducing classification bias and variance. Bias is defined as a systematic error of a learning algorithm and is impacted by the algorithm itself (Rosales-Pérez, Escalante, Gonzalez, Reyes-Garcia, & Coello, 2013). Variance describes random errors, which are caused by the uncertainty of training data or learning algorithm settings. Generally, low bias models may cause overfitting issues, which limit the capability of models to classify new instances, while low variance models may suffer from underfitting problems, which may cause them to lose their accuracy. Through bias and variance analyses, researchers proposed different methods to decompose and reduce classification errors (Neville & Jensen, 2008; Rosales-Pérez et al., 2013). Moreover, to balance the influence from bias and variance in classification tasks, ensemble learning is considered as one of the most effective strategies. Essentially, ensemble methods aggregate different algorithms together for a comprehensive decision. There were several ensemble learning paradigms proposed in the literature, such as boosting algorithms, bagging, and windowing. The basic idea of boosting is to dynamically adjust the training process to focus more on those cases that caused errors. To achieve the aim, each training sample is assigned a weight. The whole training process tends to concentrate on “hard” instances through updating weights accordingly. The boosting algorithm combines many base classifiers into a single final classifier through weights, which tends to reduce bias (Schapire & Freund, 2012). An SVM-based boosting method is also discussed in the literature. Li, Wang, and Sung (2008) implemented an RBF-based SVM in adaptive boosting (AdaBoost) algorithm and proposed AdaBoostSVM. The model is compared with DT-based and NN-based boosting on several standard datasets. Their results show that the AdaBoostSVM model performs better than other AdaBoost models with base classifiers DTs and NNs. However, studies showed that AdaBoost with strong base classifiers is not viable (Wickramaratna, Holden, & Buxton, 2001). To embed a RBF-based SVM in AdaBoost, the researchers weaken the SVM model by adaptively adjusting the kernel parameter σ when boosting proceeds (Li et al., 2008). Even though their experimental results showed that AdaBoostSVM could obtain a good performance, the AdaBoostSVM structure did not fully utilize the capability of SVM, which is a sacrifice for SVM. Another widely used ensemble paradigm is bagging, which utilizes a hybrid bootstrap sampling technique and aggregates the choices from various base classifiers. Bagging models have an advantage to reduce variance. Breiman (1996) showed that bagging can reduce classification errors by 20% on average; however, Breiman also showed that

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001

bagging could reinforce good classifiers, but make poor classifiers worse.

In the ensemble learning process, several high performance ensemble models have been studied in literature. Especially, Receiver Operating Characteristics (ROC) curve was applied as a comprehensive classifier performance criterion to replace accuracy in ensemble learning tasks (Gao, Lee, & Lim, 2006; Levesque, Durand, Gagne, & Sabourin, 2012). Studies have shown that the Area under the ROC (AUC)-based learning process can refine the traditional model parameter tuning procedure (Gao & Sun, 2007). Gao et al. (2006) proposed an ensemble maximal figure-of-merit (E-MFoM) model based on ROC optimization. The E-MFoM model integrated a statistical sampling technique to optimize a particular operating point on the ROC curve of each linear discriminant function. Their results show that E-MFoM learning achieves better classification results compared to state-of-the-art ROC optimization models. As a modification, Levesque et al. (2012) proposed a multi-objective evolutionary optimization method to train ensemble learning. In their method, false positive and true positive rates were formulated as two objectives. Then, a Non-Dominated Sorting Genetic Algorithm II (NSGA-II) was applied in the training procedure. Their results were compared with the best single classifiers in the literature and several ensemble structures by applying commonly used base classifiers, such as SVM, linear discriminant, and expression trees. Their results show that the multi-objective mechanism can obtain better performance compared to previous approaches in the literature. The research in literature indicates that ROC curves can be applied as a more effective criteria to evaluate classifier performance compared to traditional single measures, such as accuracy. However, those studies focused mainly on applying ROC curves in the model parameter tuning process and did not take into consideration how to combine the decisions from existing tuned base classifiers. In addition, evaluating ROC curves for objective functions in each training step requires a high cost of computation and memory.

As described previously, a single SVM, as one of the low bias classifiers, has been well studied for breast cancer diagnosis. SVMs have demonstrated their superior capability in breast cancer diagnosis. ROC curves have shown their abilities to optimize the parameter tuning process. However, limited research has been focused on using the ensemble learning approach to further improve SVM-based breast cancer diagnosis. Based on the nature of ensemble learning, there is room for overcoming the limitation of an individual SVM due to its singular structure type and variance.

3. Methodology

In this section, the proposed SVM-based ensemble algorithm is explained in detail. There are two main components in the proposed methodology: individual SVM classifiers and ensemble methods. Regarding various basic kernels of SVMs, the performance of each base model is investigated to show the impact of model parameters on model accuracy. Based on the individual SVM's performance, the Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) approach is proposed to leverage the strength and compromise the weakness of individual SVMs. The effectiveness of WAUCE is further compared with other ensemble strategies.

3.1. Kernel-based SVMs

Initially, SVMs are formulated as quadratic optimization problems, which can solve only linear separable classification cases. Using Lagrange multipliers, an inner product space is introduced into the dual form of the optimization model, which can extend linear to nonlinear SVM via feature mapping by kernel functions. Suppose Y is the binary output set, i.e., $\Upsilon = \{-1, +1\}$, χ^n is a n-dimensional feature space, and τ is the training set size. Given a training dataset $D = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \chi^n, y_i \in \Upsilon, i = 1, \ldots, \tau\}$, the dual form of a C-SVM model is given by

$$C-\text{SVM}: \quad \max_{\alpha} \quad -\frac{1}{2}\sum_{i,j=1}^{\tau}\alpha_i\alpha_j y_i y_j \kappa(\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{\tau}\alpha_i \tag{1}$$

$$s.t. \quad 0 \le \alpha_i \le C \quad \forall i \tag{2}$$

$$\sum_{i=1}^{\tau}\alpha_i y_i = 0 \tag{3}$$

where α_i denotes Lagrange multipliers, κ represents kernel function, and parameter C is a regularization term, which is used to balance structural and empiric risks (Cawley, 2001). Eqs. (1)–(3) are also called C-SVM, which can tolerate input noise in a dataset by adjusting parameter C. Another SVM model used in this research is ν-SVM, as shown in Eqs. (4)–(7), where $\nu \in [0, 1]$ controls the upper bound on the fraction of margin error, and it also determines the lower bound of the fraction of support vectors (Schölkopf, Smola, Williamson, & Bartlett, 2000). By adjusting parameter ν, the ν-SVM model gives another aspect of margin error control (Chen, Lin, & Schölkopf, 2005).

$$\nu-\text{SVM}: \quad \max_{\alpha} \quad -\frac{1}{2}\sum_{i,j=1}^{\tau}\alpha_i\alpha_j y_i y_j \kappa(\mathbf{x}_i \cdot \mathbf{x}_j) \tag{4}$$

$$s.t. \quad 0 \le \alpha_i \le \frac{1}{\tau} \quad \forall i \tag{5}$$

$$\sum_{i=1}^{\tau}\alpha_i y_i = 0 \tag{6}$$

$$\sum_{i=1}^{\tau}\alpha_i \ge \nu \tag{7}$$

By applying an inner product between any two points in a given feature space, χ^n, and transferring features to a higher dimensional space, a kernel method can linearly separate highly intermeshed overlapping data points in the new space. Feature mapping is achieved by using kernel functions, as follows:

$$(\mathbf{x}_i, \mathbf{x}_j) \longmapsto \kappa(\mathbf{x}_i, \mathbf{x}_j), \tag{8}$$

by satisfying

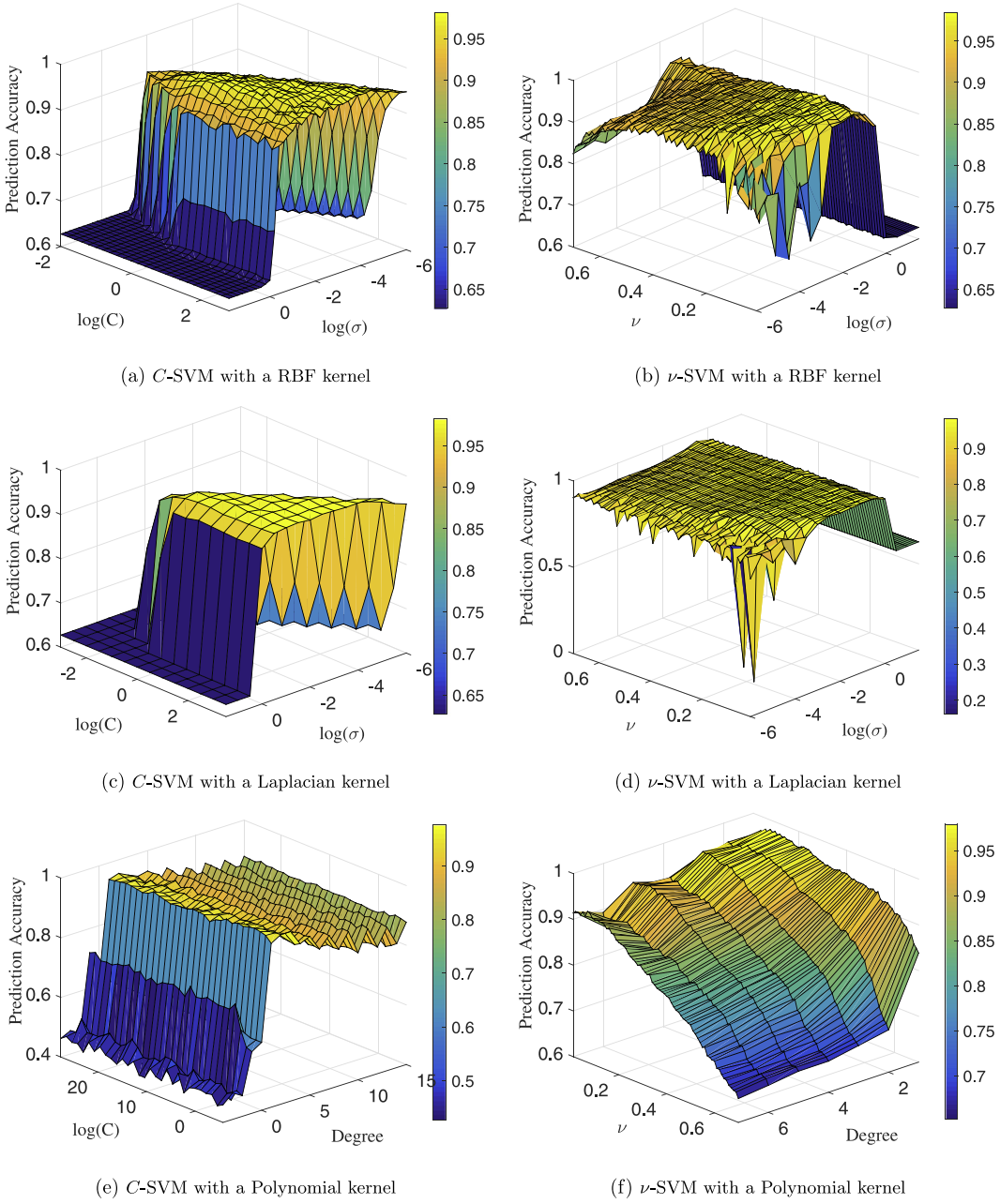
$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle, \tag{9}$$

where $\Phi(\mathbf{x})$ denotes the mapping function.

Commonly used kernel functions include the linear kernel, the polynomial kernel, and the Gaussian radial basis function (RBF) kernel, etc. This research also includes the Laplacian kernel, the Hyperbolic tangent (HT) kernel, and the ANOVA RBF kernel. The kernel functions and their commonly used default parameter settings are provided in Table 1.

Fig. 2 shows the accuracy surfaces for the C-SVM and ν-SVM models based on three kernel functions using the WDBC dataset by 10-fold cross-validation. The details of the dataset are illustrated in Table 4. Fig. 2(a) and 2(c) show that the classification accuracy of the C-SVM is reduced when the regularization term (C) becomes extremely small. On the contrary, the kernel parameter (σ) plays a more important role than the regularization term for a relatively large regularization value. For the polynomial kernel, degree (r) has greater impact on model accuracy than C because the surface is

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001



(a) C-SVM with a RBF kernel

(b) ν-SVM with a RBF kernel

(c) C-SVM with a Laplacian kernel

(d) ν-SVM with a Laplacian kernel

(e) C-SVM with a Polynomial kernel

(f) ν-SVM with a Polynomial kernel

Fig. 2. Classification accuracy surfaces for C-SVM and ν-SVM models.

Table 1
Kernel functions and default parameters for model selection.

Kernel type	Functions	Default parameters
RBF	$\kappa_r(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sigma \|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2)$	$\sigma = 1$
Polynomial	$\kappa_p(\mathbf{x}_i, \mathbf{x}_j) = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1]^r$	$r = 1$
Linear	$\kappa_l(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$	/
Laplacian	$\kappa_{lap}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sigma \|\|\mathbf{x}_i - \mathbf{x}_j\|\|)$	$\sigma = 1$
HT	$\kappa_t(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + b)$	$b = 1$
ANOVA RBF	$\kappa_a(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{k=1}^{s} \exp(-\sigma (\mathbf{x}_i^k - \mathbf{x}_j^k)^2))^d$	$\sigma = 1, d = 1$

almost parallel along $\log(C)$ axis, as shown in Fig. 2(e). Regarding the ν-SVM, the accuracy surfaces have different representations. The prediction accuracy fluctuates significantly, when kernel parameter σ becomes small for RBF- and Laplacian-based kernels, as shown in Fig. 2(b) and 2(d). In particular, the ν-SVM has relatively high accuracy on the overall parameter region for the Laplacian kernel, but there are some potentially risky points in the high accuracy region, as shown in Fig. 2(d). The Polynomial kernel-based ν-SVM has a much smoother accuracy surface, and no very flat and steep region appearance, as shown in Fig. 2(d). Moreover, ν-SVM accuracy tends to decrease while either increasing degree or ν under the Polynomial kernel.

SVM accuracy surfaces in Fig. 2 reveal the fact that various parameter settings in SVM models can significantly affect the classification accuracy. Particularly, the classification accuracy varies due to the different options of kernel functions and the structure of SVMs. The experimental results indicate that the two types of SVMs achieve different best classification accuracy under different parameter settings. In other words, the individual SVM may have its upper limit of accuracy due to the nature of the SVM structure. Potential improvement can be achieved by aggregating vari-

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001

Table 2
Confusion matrix.

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

ous SVM models together to reduce the weakness of an individual SVM, especially due to the singular SVM structure.

3.2. Ensemble SVM learning

Two SVM paradigms with six kernel functions are combined in the proposed SVM-based model to increase the diversities of base models in the ensemble structure. For the purpose of comparisons, ensemble SVM learning models are also implemented using five different fusion mechanisms for further validation of the effectiveness of WAUCE.

3.2.1. Area Under the Receiver Operating Characteristic Curve (AUC)

A confusion matrix is one of the common approaches to measure performance for classification models. In a confusion matrix, the two classes are identified as positive class $(+1)$ and negative class (-1) . As shown in Table 2, each predicted class is compared with its actual class for each instance to calculate four metrics:

- True Positives (TP) – the number of positive instances that is correctly classified as positive classes.
- False Positives (FP) – the number of negative instances that is incorrectly classified as positive classes.
- True Negatives (TN) – the number of negative instances that is correctly classified as negative classes.
- False negatives (FN) – the number of positive instances that is incorrectly classified as negative classes.

Based on the confusion matrix, other performance measures can be derived as follows:

$$\text{Error:} \quad e = \frac{FN + FP}{TN + TP + FN + FP}, \tag{10}$$

$$\text{Accuracy:} \quad a = \frac{TN + TP}{TN + TP + FN + FP} = 1 - e, \tag{11}$$

$$\text{Sensitivity:} \quad r = \frac{TP}{TP + FN}, \tag{12}$$

$$\text{Specificity :} \quad s = \frac{TN}{FP + TN}. \tag{13}$$

Based on Eq. (10), it is obvious that a classification model error is shown from only FN and FP. Given two classes, i.e., benign (B) and malignant (M), the FN and FP rates are impacted by setting up different threshold values as shown in Fig. 3(a), where the horizontal axis is a feature measure, such as the number of abnormal cells. By moving the decision threshold, the FP and FN areas keep changing, and it is impossible to minimize FN and FP rates simultaneously. However, the TP, FP, TN, and FN measures can be collected to construct a plot, which is a Receiver Operating Characteristic (ROC) curve, to show the tradeoff of FN and FP rates to model classification errors. As shown in Fig. 3(b), ROC curves are typically plotted using FP rate vs. TP rate (Bradley, 1997). Based on the ROC curve, AUC can be calculated. Suppose $1 - s$ and r are the probabilities of FP and TP, respectively. Then, AUC can be estimated by trapezoidal integration, which is expressed as Eq. (14) (Bradley, 1997)

$$\text{AUC} = \sum_{\gamma} \left\{ [r_\gamma \cdot \Delta(1-s)] + \frac{1}{2}[\Delta r \cdot \Delta(1-s)] \right\} \tag{14}$$

where $\Delta(1-s) = (1-s)_\gamma - (1-s)_{\gamma-1}$, $\Delta r = r_\gamma - r_{\gamma-1}$, and γ is an index. In practical applications, the distribution of features for different classes can vary in different shapes. Fig. 3(c) shows an example where the distribution of the average number of concave points of the cell nuclei contour is unimodal for benign cases but bimodal for malignant cases in digitized images of a fine needle aspirate (Mangasarian et al., 1995). If a threshold-based diagnostic model is built based on the average number of concave points, the ROC, as shown in Fig. 3(d), can be obtained by varying the model configurations. Based on trapezoidal integration, AUC of the threshold-based model is approximately 78.59%, which is a comprehensive evaluation of model structure. Obviously, AUC is an ensemble evaluation of a classification model performance, which can provide more information than a single accuracy measurement. In this research, the AUC information is used in the ensemble weighting process to enhance breast cancer diagnosis.

3.2.2. Weighted Area Under the ROC Curve Ensemble (WAUCE)

Given an input feature vector $\mathbf{x} \in \chi^n$, a classification result $h_\eta(\mathbf{x}) \in \Upsilon$ is obtained based on each base classifier $h_\eta \in B$, where B is the base classifier set $B = \{h_\eta : \eta = 1, \ldots, m\}$, and m is the number of base classifiers. The aim of the ensemble learning is to create an enhanced composite classifier, $H(x)$, through amalgamating diagnoses from the different base SVM models into one single decision. The structure of the ensemble learning algorithm is presented as Algorithm 1. A sampling with replacement approach is used here to construct different base classifiers. To train a base classifier, a subsample is selected from the original dataset using the sampling with replacement approach in each iteration. $F(\mathbf{x}, h)$ in Algorithm 1 represents a fusion strategy, which is used in ensemble learning to aggregate the decisions from different classifiers. Major voting is a typical fusion strategy in many studies. However, majority voting considers the decision from each classifier equally, and neglects the influence from those low accuracy classifiers. To overcome the shortcoming of majority voting, weighted fusion approaches were also applied. The typical format of weighted fusion is given as Eq. (15).

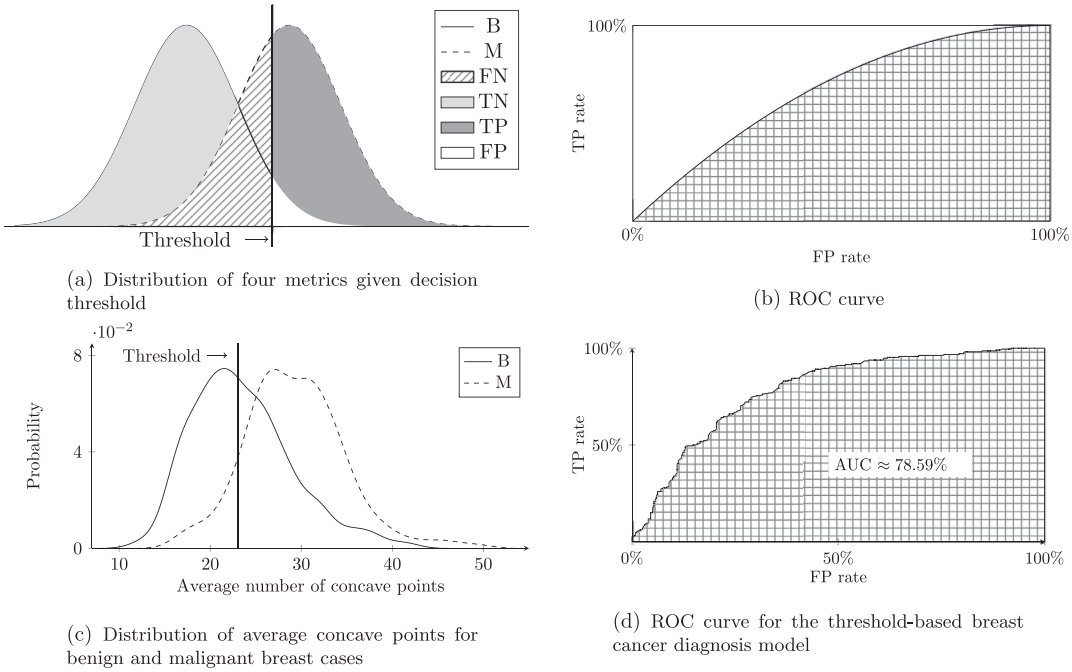
Algorithm 1 Weighted Area Under the ROC Curve Ensemble (WAUCE).

```
Training dataset D = {(x_i, y_i) : x_i ∈ χ^n, y_i ∈ Υ, i = 1, ..., τ}
B = {h_η : η = 1, ..., m}
F(x, h)
procedure ENSEMBLE(D, B, F)
    for η = 1, ..., m do
        S_η = Sampling(D)
        h_η ← h_η(S_η)
    H(x) = F(x, h_1, ..., h_m)
    return H(x)
```

$$H(\mathbf{x}) = F\left[\sum_{\eta=1}^{m} w_\eta h_\eta(\mathbf{x})\right] \tag{15}$$

where w_η is the weight for each base classifier. In the proposed model, the AUC is used as w_η for the SVM ensemble, i.e., WAUCE. In particular, five commonly used additional fusion strategies are provided to compare the effectiveness of the WAUCE model.

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001



(a) Distribution of four metrics given decision threshold

(b) ROC curve

(c) Distribution of average concave points for benign and malignant breast cases

(d) ROC curve for the threshold-based breast cancer diagnosis model

Fig. 3. ROC curve and AUC.

- Weighted Area Under the ROC Curve Ensemble (WAUCE)
The AUC performance metric is used to weigh different diagnosis results from base classifiers in the WAUCE model. The strategy can be expressed mathematically as Eq. (16).

$$H(\mathbf{x}) = \begin{cases} -1 & \sum_{\eta=1}^{m} w_\eta h_\eta(\mathbf{x}) < 0 \\ +1 & \text{Otherwise} \end{cases} \tag{16}$$

where $w_\eta = \frac{\text{AUC}_\eta}{\sum_{j=1}^{m} \text{AUC}_j}, \forall \eta$.

- Majority Voting Ensemble (MVE)
MVE is also called plurality vote (PV) or the basic ensemble method (BEM), which is widely used in many applications (Rokach, 2010). The MVE can be expressed as Eq. (17).

$$H(\mathbf{x}) = \arg\max_{\forall c \in \Upsilon} \left\{ \sum_{\eta=1}^{m} I[h_\eta(\mathbf{x}), c] \right\} \tag{17}$$

where $h_\eta(\mathbf{x})$ is the classification result for the ηth base classifier, and $I(h, c)$ denotes an indicator function, which is defined as Eq. (18) in this paper.

$$I(h, c) = \begin{cases} +1 & h = c \\ -1 & h \neq c \end{cases} \tag{18}$$

- Maximum Ensemble (MaxE)
MaxE always selects the largest decision value from the base model set to make the final decision, which can be expressed mathematically as Eq. (19).

$$H(\mathbf{x}) = \max\{h_\eta(\mathbf{x}) : \eta = 1, \ldots, m\} \tag{19}$$

- Minimum Ensemble (MinE)
Opposite to MaxE, the MinE always applies the smallest decision value in the base model set to provide the final decision, which can be expressed mathematically as Eq. (20).

$$H(\mathbf{x}) = \min\{h_\eta(\mathbf{x}) : \eta = 1, \ldots, m\} \tag{20}$$

- Weighted Accuracy Ensemble (WAE)
The final decision for WAE is based on a comprehensive evaluation of all the base model decision results through accuracy-based weighting. Training accuracies are used to weigh different diagnosis results. The strategy can be expressed mathematically as Eq. (21).

$$H(\mathbf{x}) = \begin{cases} -1 & \sum_{\eta=1}^{m} w_\eta h_\eta(\mathbf{x}) < 0 \\ +1 & \text{Otherwise} \end{cases} \tag{21}$$

where $w_\eta = \frac{a_\eta}{\sum_{j=1}^{m} a_j}, \forall \eta,$ and a_η is the training accuracy of the ηth base classifier.

- Single Best (SB)
In addition, the five fusion strategies are compared with the best individual base SVM model during each experimental test. For the SB model, the final decision is made based on the base model with the highest training accuracy, which can be expressed mathematically as Eq. (22).

$$H(\mathbf{x}) = h_\Delta(\mathbf{x}), \Delta = \arg\max_\eta \{a_\eta : \eta = 1, \ldots, m\} \tag{22}$$

4. Experiment results and analysis

To further evaluate the proposed model for generalization, several experiments are conducted and analyzed in detail to explore the actual improvements.

4.1. Data description

To test the effectiveness of the proposed ensemble SVM structures for breast cancer diagnosis, two standard breast cancer datasets are applied. The two datasets were collected from the University of Wisconsin Hospitals, Madison by Mangasarian et al. (1995). In addition, to demonstrate how the proposed WAUCE method performs on practical large scale, modern datasets, a latest version SEER breast cancer dataset (SEER, 2017) is also tested in this section. A brief description of each dataset is given as follows:

- Wisconsin Original Breast Cancer (WBC) Dataset

The WBC dataset includes 699 observations (65.52% benign, 34.48% malignant). Sixteen instances that include missing values for attribute bare nuclei are removed from the dataset during data preprocessing. The distribution of each feature is summarized as Table 3.

- Wisconsin Diagnostic Breast Cancer (WDBC) Dataset
  There are 569 instances (62.74% benign, 37.26% malignant) and 32 patient attributes, which include a patient ID, 30 tumor features, and one class indicator in the WDBC dataset. Tumor features were collected from 10 aspects: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These features were collected from a digitized image of a fine needle aspirate (FNA) of a breast mass. The mean, standard error, and "worst" or "largest" of these features were computed for each image, which resulted in a total of 30 features. A summary of the attributes is given in Table 4.
- SEER Breast Cancer Dataset
  This is a dataset collected by National Cancer Institutes Surveillance, Epidemiology, and End Results (SEER) program. The program is responsible for collecting incidence and survival data related to cancer at nine anatomical sites, e.g., breast, colon and rectum, urinary tract. The dataset is publicly available and can be accessed by signing a research data agreement. The latest version, submitted in November 2016, includes more than 9.6 million cancer instances from 1973 to 2014. The breast cancer dataset, which includes 800,000 instances, is used in this study (SEER, 2017). In the dataset, there are 133 variables to record the patient information from different perspectives, such as geographic area, race, and stage of cancer.
  In this research, 14 features are used as predictors (Delen et al., 2005). However, some of the variables applied to describe patient statuses contain many missing values, such as tumor size and number of lymph nodes examined, due to coding changes. To avoid the impact of coding system changes, only the instances with the diagnosis year after 2004 are used. In addition, the logic applied to determine survivability of breast cancer patients from the SEER dataset is based on three variables: survival months, vital status recode, which describes whether or not the patient is alive as of the cut-off date, and cause of death. Those patients who survive for 60 months (five years) and are still alive after diagnosis are defined as survived. For those cases with the surviving months of fewer than 60 months and the cause of death as breast cancer, they are defined as not survived (Kate & Nadig, 2017). Given that survivability is defined as surviving for 60 months after diagnosis, the instances that were diagnosed as fewer than 60 months from the latest year of submission are excluded. After the data cleansing and data preparation, 82,707 instances, which include 76,716 positive and 5991 negative instances, are obtained. A summary of the variables is given in Table 5. To avoid the influence of dataset unbalance, a sampling without replacement approach is used to randomly undersample the positive class.

Table 3
Summary of attributes for WBC.

Attributes	Domain	Mean	SD
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

Table 4
Range of each attributes in WDBC.

Attributes	Range		
	Mean	Standard error	Largest value
Radius	6.98–28.11	0.11–2.87	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	185.20–4254.00
Smoothness	0.05–0.16	0.00–0.03	0.07–0.22
Compactness	0.02–0.35	0.00–0.14	0.03–1.06
Concavity	0.00–0.43	0.00–0.40	0.00–1.25
Concave points	0.00–0.20	0.00–0.05	0.00–0.29
Symmetry	0.11–0.30	0.01–0.08	0.16–0.66
Fractal dimension	0.05–0.10	0.00–0.03	0.06–0.21

Table 5
Range of each attributes for the preprocessed SEER breast cancer dataset.

Categorical Attributes	Number of unique values		
Race	28		
Marital status	6		
Primary site code	9		
Histology	88		
Behavior	2		
Grade	5		
Extension of disease	27		
Lymph node involvement	7		
Stage of cancer	4		
Site specific surgery code	46		
Continuous Attributes	**Mean**	**SD**	**Range**
Age	58.38	12.68	13–98
Tumor size	20.76	18.12	0–200
Number of positive nodes	1.08	3.00	0–84
Number of nodes	6.96	6.98	1–90

4.2. Design of experiments

During the tests, a k-fold cross-validation approach is applied to estimate the generalization error for constructed models (Levesque et al., 2012). The k-fold cross-validation includes K iterations. The whole dataset is split into K roughly equal-sized partitions. In the kth iteration, a model is trained based on remaining $K-1$ partitions and tested using the kth partition to get testing performance P_k. The overall performance in k-fold cross-validation is averaged based on these results from K iterations, as shown in Eq. (23).

$$\bar{P} = \frac{1}{K}\sum_{k=1}^{K} P_k \tag{23}$$

where P_k is a performance metric from a, r, s, or AUC for kth partition.

In addition, to compare the reliability of the ensemble models, variance is measured based on standard deviation (σ), which is defined as Eq. (24).

$$\sigma(P) = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}(\bar{P}_l - \bar{\bar{P}})^2} \tag{24}$$

where L is the number of replications.

In the experimental test, 12 SVMs ($m = 12$, i.e., C-SVM and ν-SVM with six kernel functions, respectively) are included together in the ensemble process. The kernel related parameters of each base model are set as default as given in Table 1. In the experimental tests, C and ν are set as 1 and 0.2 for C-SVM and ν-SVM, respectively. Ten-fold cross-validation ($K = 10$) is repeated with five replications ($L = 5$) for each model. The performance of the WAUCE model is compared with the other five SVM-based

Table 6
Diagnosis accuracy and model training time comparisons on the WBC dataset.

Models	Accuracy (a) (%)	Specificity (s) (%)	Sensitivity (r) (%)	AUC (%)	Training time (seconds) (SD)
MVE	96.86	96.92	96.70	96.81	0.46 (0.014)
MaxE	94.51	92.06	**99.36**	95.71	0.46 (0.005)
MinE	94.79	**97.87**	89.15	93.51	0.45 (0.006)
WAE	**97.08**	97.16	97.06	**97.11**	0.50 (0.008)
SB	96.62	97.18	95.74	96.46	0.50 (0.006)
AdaBoost	96.09	97.05	94.40	95.73	0.39 (0.013)
BCT	95.59	96.55	93.85	95.20	0.35 (0.006)
WAUCE	**97.10**	**97.23**	**97.11**	**97.17**	0.53 (0.005)

Table 7
Diagnosis accuracy and model training time comparisons on the WDBC dataset.

Models	Accuracy (a) (%)	Specificity (s) (%)	Sensitivity (r) (%)	AUC (%)	Training time (seconds) (SD)
MVE	97.15	99.39	93.32	96.36	0.76 (0.016)
MaxE	89.65	84.19	**98.63**	91.41	0.82 (0.013)
MinE	89.64	**99.95**	72.67	86.31	0.82 (0.009)
WAE	**97.54**	99.43	94.49	**96.96**	0.92 (0.006)
SB	96.77	98.64	93.59	96.12	0.82 (0.014)
AdaBoost	95.86	97.35	93.46	95.40	0.48 (0.004)
BCT	94.00	96.10	90.54	93.32	0.49 (0.020)
WAUCE	**97.68**	**99.49**	**94.75**	**97.12**	0.87 (0.040)

Table 8
Diagnosis accuracy and model training time comparisons on the SEER dataset.

Models	Accuracy (a) (%)	Specificity (s) (%)	Sensitivity (r) (%)	AUC (%)	Training time (seconds) (SD)
MVE	75.56	**77.53**	73.59	75.56	292.91 (1.673)
MaxE	50.12	0.25	**99.97**	50.11	304.72 (1.754)
MinE	50.08	**99.97**	0.18	50.08	294.35 (2.793)
WAE	**76.30**	72.65	79.93	**76.29**	355.78 (14.405)
SB	57.31	45.90	68.91	57.40	356.44 (16.803)
AdaBoost	75.54	75.52	75.62	75.57	3.82 (0.050)
BCT	75.54	75.52	75.62	75.57	3.01 (0.106)
WAUCE	**76.42**	72.80	**80.02**	**76.41**	355.78 (14.404)

ensembles. In addition, two typical ensemble models, AdaBoost (Freund, Schapire et al., 1996) and Bagging Classification Trees (BCT) (Breiman, 1996) are also tested to compare with the WAUCE model under the same experiment conditions. The number of base trees in AdaBoost and BCT is set as 12, respectively. Performance of each model is estimated using WBC, WDBC, and SEER datasets without feature selection and feature extraction. In the experimental tests, R packages are used for the implementation of base SVM models, and the ensemble structure is coded in R language. All the experimental tests are performed on a 64-bit Windows 10 with Intel i7 processor (4.00 gigahertz) and 16 gigabyte random-access memory (RAM).

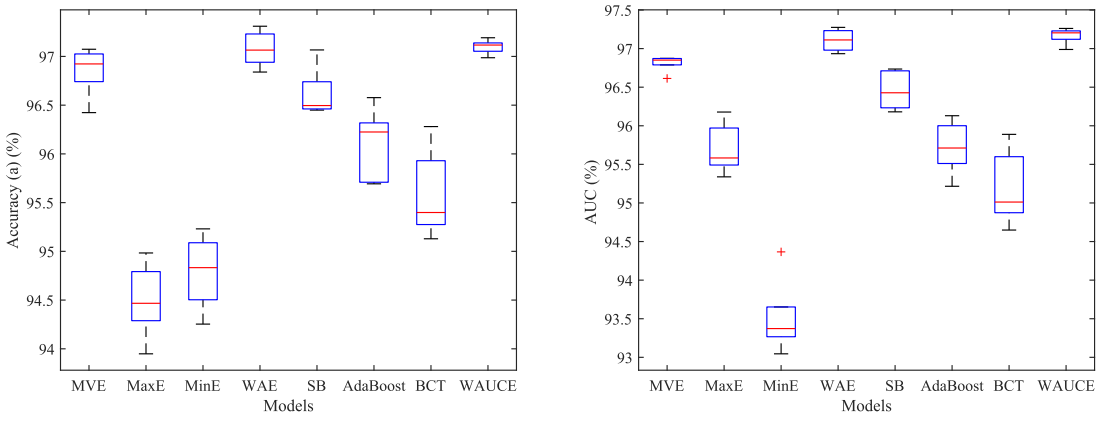
4.3. Experimental results

The performance of the proposed model is evaluated from three aspects: effectiveness improvement; reliability improvement; and the improvement over existing works.

- Analysis of diagnosis accuracy and training efficiency
  During each test, the accuracy (a), specificity (s), sensitivity (r), and AUC for WBC, WDBC and SEER datasets are collected and summarized in Tables 6–8 for the eight models, respectively. The top two highest values are highlighted in bold for each performance measure in each table. To compare training efficiency, training times are also provided in the last column of each table.
  For small size datasets, i.e., WBC and WDBC, the comparison results of each model with the proposed model show that the proposed WAUCE model structure always outperforms others in terms of accuracy and AUC. The best average accuracy achieved by the WAUCE model is 97.68% on the WDBC dataset. In particular, SB represents the best performance level that individual SVM classifiers can achieve, and WAUCE outperforms SB for all the performance metrics on both datasets. Especially, WAUCE, as well as the basic MVE, is better than BCT for all the performance metrics on both datasets, which indicates that the proposed SVM-based bagging is better than DT-based bagging, which also supports the initial objective of this research to improve the ensemble process using high performance base classifiers. MaxE and MinE outperform other models in terms of specificity or sensitivity due to their biased diagnosis. For instance, MaxE improves the sensitivity while sacrificing the specificity significantly. In other words, it gives many false positive diagnoses of breast cancer in the testing experiments. The WAUCE model balances the performance and obtains high performance on both sensitivity and specificity. In short, WAUCE outperforms other methods in terms of accuracy and AUC measures. It implies that the AUC-based weight strategy can effectively improve the breast cancer diagnosis accuracy.
  For the SEER breast cancer dataset, the proposed WAUCE model obtains the highest accuracy and AUC, which indicates the ability of using WAUCE for practical large data size breast cancer diagnosis. MinE manifests the highest specificity, i.e., 99.97%, to diagnose negative cases correctly, but can diagnose only positive cases with sensitivity 0.18%, which indicates its seriously biased diagnostic performance for the large dataset. In practice, the MinE model may miss numerous malignant breast cancer cases due to its low sensitivity. A similar result can also be found for the MaxE model. Overall, the best average accuracy is achieved as 76.42% by the WAUCE model, which is 33.34% improvement compared to the single best model.
  In terms of training efficiency, DT-based ensemble methods, i.e., AdaBoost and BCT, require the least training computation time, as shown in Tables 6–8. WAUCE is not as efficient as other models in terms of training efficiency on small datasets based on the results in Tables 6 and 7. However, for the SEER dataset, WAUCE achieves a similar computation time level as WAE and SB methods, as shown in Table 8.
- Analysis of diagnosis variance
  One of the critical challenges for disease diagnosis is the model reliability, which indicates the model not only can achieve a high diagnosis accuracy, but also can sustain the high accuracy stably. The reliability improvement of the ensemble technique can be investigated from the performance variance. Based on Eq. (24), the variance of each model performance is collected, as provided by Tables 9–11, for all the relevant performance metrics. Interestingly, MVE is not better than either AdaBoost or BCT particularly for WDBC and SEER datasets, but the AUC-based weighted process improves the ensemble learning and makes the WAUCE model significantly outperform both AdaBoost and BCT in terms of accuracy variance reduction. The

Table 9
Diagnosis variance comparisons on the WBC dataset.

Models	$\sigma(a)$ (%)	$\sigma(s)$ (%)	$\sigma(r)$ (%)	σ(AUC) (%)
MVE	0.39	0.50	0.21	0.33
MaxE	0.39	0.50	0.21	0.33
MinE	0.38	0.33	1.01	0.50
WAE	0.19	0.29	0.18	0.14
SB	0.26	0.44	0.66	0.25
AdaBoost	0.38	0.58	0.75	0.35
BCT	0.46	0.39	0.73	0.49
WAUCE	**0.08**	0.16	0.33	**0.11**



(a) Accuracy (b) AUC

Fig. 4. Diagnosis reliability comparisons on the WBC dataset.

Table 10
Diagnosis variance comparisons on the WDBC dataset.

Models	$\sigma(a)$ (%)	$\sigma(s)$ (%)	$\sigma(r)$ (%)	σ(AUC) (%)
MVE	1.00	1.62	0.30	0.88
MaxE	1.00	1.62	0.30	0.88
MinE	0.36	0.12	1.10	0.53
WAE	0.32	0.16	1.05	0.50
SB	0.51	0.36	1.20	0.59
AdaBoost	0.30	0.25	0.93	0.42
BCT	0.69	0.62	1.77	0.88
WAUCE	**0.22**	0.31	0.95	**0.35**

Table 11
Diagnosis variance comparisons on the SEER dataset.

Models	$\sigma(a)$ (%)	$\sigma(s)$ (%)	$\sigma(r)$ (%)	σ(AUC) (%)
MVE	0.96	3.23	5.17	0.99
MaxE	0.18	0.43	0.05	0.19
MinE	**0.09**	0.04	0.23	**0.09**
WAE	0.22	0.89	1.31	0.24
SB	5.22	23.62	19.58	5.24
AdaBoost	0.25	1.03	1.17	0.23
BCT	0.52	5.43	4.89	0.51
WAUCE	**0.11**	1.23	1.03	**0.11**

comparison results show that the WAUCE model obtains the smallest diagnosis accuracy variance for both WBC and WDBC datasets. In particular, the proposed WAUCE model can achieve around 69.23% and 56.86% accuracy variance reduction in comparison to the SB classifiers on WBC and WDBC datasets, respectively. For the SEER dataset, the MinE model obtains the lowest variance on both accuracy and AUC. The reason is due to the serious bias of MinE, which has little capacity to diagnose positive cases but has the strongest capacity to diagnose negative cases. WAUCE is the second best model with low variance of both accuracy and AUC on the SEER dataset. Compared to the SB model, WAUCE can obtain 97.89% variance reduction.

To further investigate the reliability improvement of the proposed WAUCE model, the comparison is also conducted using graphical analysis. The variance of accuracy and AUC are illustrated through box plots for the two datasets in Figs. 4–6. Comparing all the models, it is obvious that the proposed WAUCE model can maintain high accuracies with small diagnosis variance, which also confirms our initial objective to improve the breast cancer diagnosis accuracy while still reducing the diagnosis variance.

- Statistical test

Table 12
Friedman test results.

Test item	R_{WAUCE}	χ_r^2	$F_F(2.76)$	Decision
Accuracy (a)	1.00	13.40	22.63	Positive
$\sigma(a)$	1.33	10.40	2.90	Positive
AUC	1.00	12.50	8.45	Positive
σ(AUC)	1.33	6.70	0.93	Negative

To analyze the statistical significance of the performance from all methods, a Friedman test is applied (Friedman, 1937). Based on performance ranking of different algorithms from different datasets, the Friedman test can measure the statistical difference among the algorithms. The Friedman estimator F_F, which follows a Fisher distribution, can be measured in Eq. (25).

$$F_F = \frac{(N_D - 1)\chi_r^2}{N_D(N_M - 1) - \chi_r^2} \tag{25}$$

$$\chi_r^2 = \frac{12N_D}{N_M(N_M+1)}\left(\sum_{i=1}^{N_M} R_i^2 - \frac{N_M(N_M+1)^2}{4}\right) \tag{26}$$

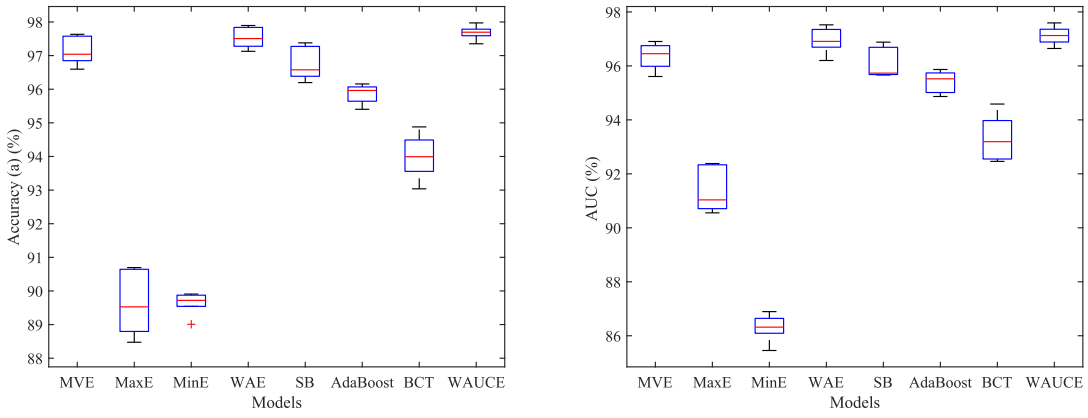
$$R_i = \frac{\sum_{j=1}^{N_D} r_{ij}}{N_D} \tag{27}$$

where N_M is the number of methods, N_D is the number of datasets compared, R_i is the average rank for the ith method, and r_{ij} denotes the rank of ith method on the jth dataset.

For each evaluation dataset, the model accuracy and AUC are ranked from one to the number of methods, respectively. In this comparison, eight models are considered. $r_{ij}^a = 1$ represents the highest accuracy or AUC, and $r_{ij}^a = 8$ indicates the worst accuracy or AUC. For model diagnosis variance test, $r_{ij}^\sigma = 1$ inducts the lowest diagnosis accuracy or AUC variation, and $r_{ij}^\sigma = 8$ indicates the model with the highest accuracy or AUC variation. In this test, F_F follows a Fisher distribution with $N_M - 1$ and $(N_M - 1)(N_D - 1)$ degrees of freedom, and the confidence level α is set as 0.05. $N_M = 8$ and $N_D = 3$, with degree of freedom $N_M - 1 = 7$ and $(N_M - 1)(N_D - 1) = 14$ are applied, which obtain a critical value of the Fisher distribution $F(7, 14) = 2.76$.

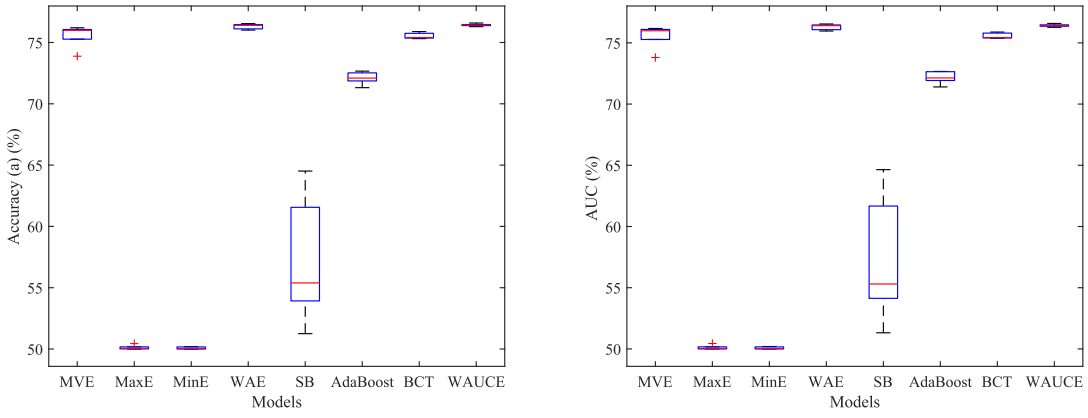
Table 12 shows the Friedman test results for accuracy (a), $\sigma(a)$, AUC, and σ(AUC). The average ranks for WAUCE (R_{WAUCE}) are the best among all the experimental models in terms of the four performance measures. In addition, the results show that the statistical difference of the rank for accuracy (a), $\sigma(a)$, and AUC are also significant.

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001



(a) Accuracy (b) AUC

Fig. 5. Diagnosis reliability comparisons on the WDBC dataset.



(a) Accuracy (b) AUC

Fig. 6. Diagnosis reliability comparisons on the SEER dataset.

Table 13
Comparison of existing and recent research model results for the WBC dataset.

Models	Accuracy (a) (%)
C4.5 decision tree	93.47
Naive Bayes	95.93
SVM-RBF kernel	96.31
SMO+J48+NB+IBk	97.28 (Salama et al., 2012)
Supervised fuzzy clustering	95.57 (Abonyi & Szeifert, 2003)
Weighted vote-based ensemble	97.42 (Bashir et al., 2015)
WAUCE (10-fold)	97.10

Table 14
Comparison of existing and recent research model results for the WDBC dataset.

Models	Accuracy (a) (%)
K-SVM	94.19
Naive Bayes	92.17
SVM-RBF kernel	96.67
RBF networks	93.70
Supervised fuzzy clustering	95.57 (Abonyi & Szeifert, 2003)
Weighted vote based ensemble	95.09 (Bashir et al., 2015)
WAUCE (10-fold)	97.68

- Comparison with existing models
  To compare the performance of the proposed WAUCE model with other algorithms in the literature, some existing methods have been implemented as a benchmark for the two standard breast cancer datasets, WBC and WDBC. The performance of the existing models are measured using 10-fold cross-validation approach with five replications. For those models, which are not publicly available, the original accuracy results from the literature are cited as shown in Tables 13 and 14.
  For the WBC dataset, Table 13 shows that WAUCE model is quite competitive, and the WAUCE model outperforms most of the other classifiers. However, due to the use of feature extraction and feature selection in Salama, Abdelhalim, and Zeid (2012), the performance for SMO+J48+NB+IBk is slightly better. Also, weighted vote-based ensemble Bashir, Qamar, and Khan (2015) achieved a comparable accuracy, because they applied feature selection, and also performed only one replication of 10-fold cross validation that could have yielded a better accuracy without consideration of variance, which is different from our study. For the implemented models, it is obviously that the proposed WAUCE model achieves the highest accuracy. Table 14 presents the comparison results for the WDBC dataset. It is obvious that the WAUCE model outperforms most of the other models, except SVM-RBF kernel (Aruna, Rajagopalan, & Nandakishore, 2011). However, the SVM-RBF kernel method has not been cross validated in Aruna et al. (2011). In this research, the proposed WAUCE model outperforms the implemented SVM-RBF kernel model where 10-fold cross-validation is performed. The proposed WAUCE model outperforms the weighted vote-based ensemble model from Bashir et al. (2015), which applies the feature selection process.

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001

4.4. Discussion

Through aggregating SVM models into the ensemble mechanism, the proposed ensemble model obtains better performance on both effectiveness and model reliability. The ensemble SVM learning model especially can reduce the diagnosis variance by 56.86–69.23%, i.e., $\frac{\sigma(a_{SB})-\sigma(a_{WAUCE})}{\sigma(a_{SB})}$; meanwhile, it increases the accuracy by 0.50%-0.94%, i.e., $\frac{a_{WAUCE}-a_{SB}}{a_{SB}}$, as compared to SB, based on WBC and WDBC datasets. For the large scale breast cancer dataset, the proposed WAUCE model can reduce the accuracy variation by around 97.89%, while it still increases accuracy by 33.34%, compared to the best single SVM model.

The best accuracies achieved by the proposed WAUCE model are 97.10%, 96.68%, and 76.42% for WBC, WDBC, and SEER datasets, respectively. The lowest accuracy variances obtained by the proposed WAUCE model are 0.08% and 0.22% for WBC and WDBC datasets, respectively. For the SEER dataset, although the proposed WAUCE model achieves only second lowest accuracy variance, the accuracy of WAUCE is far better than MinE, which has the lowest accuracy variance. Overall, the results show that the accuracy and reliability of the patient disease diagnosis process are significantly increased. Several aspects contribute to the good performance of the proposed model. Typically, bagging ensembles are formed from a single model, which is also called pure-bagging (West, Mangiameli, Rampal, & West, 2005). As a result, the weakness and drawbacks of the single model, which mainly come from the limitation of the model structure, cannot be overcome. Instead, two types of SVM structures and six different kernel functions are adopted in the WAUCE SVM model, which can not only increase the ensemble model structure diversities, but also increase the model parameter diversities. Also, the 12 SVM structures provide a class of strong bases in the WAUCE model, which significantly benefit the ensemble process. In addition, instead of using a typical majority voting approach, the weighted ensemble mechanism treats the decision from each individual classifier differently, which can increase the contribution of those good base classifiers and weaken the opinion from poor classifiers. The AUC, as a comprehensive evaluation of each individual classifier's performance, is applied especially to the decision tradeoff, which significantly promotes the decision-making process compared with other ensemble strategies.

5. Conclusions and future work

In this research, an SVM-based weighted AUC ensemble learning model is proposed for breast cancer diagnosis. *C*-SVM and ν-SVM with six kernel functions are utilized to increase the diversity of the base model set. Five fusion strategies are defined to aggregate the decisions from different base models to compare with the proposed WAUCE model. All the results are conducted using two standard breast cancer datasets and one large real dataset from the perspectives of model effectiveness and performance reliability. The results show that the proposed WAUCE model can significantly increase cancer diagnosis performance. To highlight the importance of model reliability for illness diagnosis, the performance variances are also compared in this paper. The proposed WAUCE structure can especially reduce variance of diagnosis accuracy by up to 69.23% while still increase accuracy by up to 0.94% compared to single best models on small datasets. In terms of the SEER breast cancer dataset, the proposed WAUCE model especially reduces the variance by around 97.89%, while it still increases accuracy by 33.34%, compared to the best single SVM model.

In the future, there are several aspects that can be extended based on this research. The SVM ensemble model structures can be used for other breast cancer datasets; relevant feature selection and extraction techniques can be applied to the model feature preparation process. Then, the SVM ensemble learning can also be used for other disease diagnoses, such as thyroid cancer, oral cancer, and diabetes. Also, in terms of computation time, parallel computation techniques can be helpful to accelerate the training process for the proposed WAUCE model.

References

Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters, 24*(14), 2195–2207.
Ades, F., Zardavas, D., Bozovic-Spasojevic, I., Pugliano, L., Fumagalli, D., de Azambuja, E., et al. (2014). Luminal B breast cancer: Molecular characterization, clinical management, and future perspectives. *Journal of Clinical Oncology, 32*(25), 2794–2803.
Aruna, S., Rajagopalan, S., & Nandakishore, L. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology, 2*, 37–45.
Ayat, N.-E., Cheriet, M., & Suen, C. Y. (2005). Automatic model selection for the optimization of SVM kernels. *Pattern Recognition, 38*(10), 1733–1745.
Bashir, S., Qamar, U., & Khan, F. H. (2015). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. *Quality & Quantity, 49*(5), 2061–2076.
Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.
Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
Cawley, G. C. (2001). Model selection for support vector machines via adaptive step–size tabu search. In *Artificial neural nets and genetic algorithms* (pp. 434–437). Springer.
Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning, 46*(1–3), 131–159.
Chauhan, N., Ravi, V., & Chandra, D. K. (2009). Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Systems with Applications, 36*(4), 7659–7665.
Chen, H.-L., 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.
Chen, P.-H., Lin, C.-J., & Schölkopf, B. (2005). A tutorial on ν-support vector machines. *Applied Stochastic Models in Business and Industry, 21*(2), 111–136.
Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning, 20*(3), 273–297.
Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine, 34*(2), 113–127.
Forouzanfar, M. H., Foreman, K. J., Delossantos, A. M., Lozano, R., Lopez, A. D., Murray, C. J., & Naghavi, M. (2011). Breast and cervical cancer in 187 countries between 1980 and 2010: A systematic analysis. *The Lancet, 378*(9801), 1461–1484.
Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning: 96* (pp. 148–156).
Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701.
Friedrichs, F., & Igel, C. (2005). Evolutionary tuning of multiple SVM parameters. *Neurocomputing, 64*, 107–117.
Gao, S., Lee, C.-H., & Lim, J. H. (2006). An ensemble classifier learning approach to ROC optimization. In *Proceedings of the 80th International Conference on Pattern Recognition: 2* (pp. 679–682). IEEE.
Gao, S., & Sun, Q. (2007). Improving semantic concept detection through optimizing ranking function. *IEEE Transactions on Multimedia, 9*(7), 1430–1442.
Gomes, T. A., Prudêncio, R. B. C., Soares, C., Rossi, A. L., & Carvalho, A. (2010). Combining meta-learning and search techniques to SVM parameter selection. In *Proceedings of the 11th Brazilian Symposium on Neural Networks* (pp. 79–84). IEEE.
Gupta, S., Kumar, D., & Sharma, A. (2011). Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering (IJCSE), 2*(2), 188–195.
Ishikawa, T., Takahashi, J., Takemura, H., Mizoguchi, H., & Kuwata, T. (2014). Gastric lymph node cancer detection using multiple features support vector machine for pathology diagnosis support system. In *Proceedings of the 15th International Conference on Biomedical Engineering* (pp. 120–123). Springer.
Kamruzzaman, J., & Begg, R. K. (2006). Support vector machines and other pattern recognition approaches to the diagnosis of cerebral palsy gait. *IEEE Transactions on Biomedical Engineering, 53*(12), 2479–2490.
Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications, 36*(2), 3465–3469.
Kate, R. J., & Nadig, R. (2017). Stage-specific predictive models for breast cancer survivability. *International Journal of Medical Informatics, 97*, 304–311.
Khan, M. U., Choi, J. P., Shin, H., & Kim, M. (2008). Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *Proceedings of the EMBS 30th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society* (pp. 5148–5151). IEEE.

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001

Kim, W., Kim, K. S., Lee, J. E., Noh, D.-Y., Kim, S.-W., Jung, Y. S., et al. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of Breast Cancer, 15*(2), 230–238.
Levesque, J.-C., Durand, A., Gagne, C., & Sabourin, R. (2012). Multi-objective evolutionary optimization for generating ensembles of classifiers in the ROC space. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation* (pp. 879–886). ACM.
Li, X., Wang, L., & Sung, E. (2008). Adaboost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence, 21*(5), 785–795.
Liu, E. T., & Sotiriou, C. (2002). Defining the galaxy of gene expression in breast cancer. *Breast Cancer Research, 4*(4), 141.
Liu, H., Zhang, R., Luan, F., Yao, X., Liu, M., Hu, Z., & Fan, B. T. (2003). Diagnosing breast cancer based on support vector machines. *Journal of Chemical Information and Computer Sciences, 43*(3), 900–907.
Lorena, A. C., & De Carvalho, A. C. (2008). Evolutionary tuning of SVM parameter values in multiclass problems. *Neurocomputing, 71*(16), 3326–3334.
Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research, 43*(4), 570–577.
Naveen, N., Ravi, V., Rao, C. R., & Chauhan, N. (2010). Differential evolution trained radial basis function network: application to bankruptcy prediction in banks. *International Journal of Bio-Inspired Computation, 2*(3–4), 222–232.
Neville, J., & Jensen, D. (2008). A bias/variance decomposition for models using collective inference. *Machine Learning, 73*(1), 87–106.
Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research, 4*, 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.
Ravi, V., Reddy, P., & Zimmermann, H.-J. (2000). Pattern classification with principal component analysis and fuzzy rule bases. *European Journal of Operational Research, 126*(3), 526–533.
Ravi, V., & Zimmermann, H.-J. (2000). Fuzzy rule based classification with feature selector and modified threshold accepting. *European Journal of Operational Research, 123*(1), 16–28.
Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1–2), 1–39.
Rosales-Pérez, A., Escalante, H. J., Gonzalez, J. A., Reyes-Garcia, C. A., & Coello, C. A. C. (2013). Bias and variance multi-objective optimization for support vector machines model selection. In *Pattern recognition and image analysis* (pp. 108–116). Springer.
Salama, G. I., Abdelhalim, M., & Zeid, M. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technology, 1*(1), 0764–2277.
Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and algorithms*. MIT Press.
Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation, 12*(5), 1207–1245.
Shah, V., Turkbey, B., Mani, H., Pang, Y., Pohida, T., Merino, M. J., et al. (2012). Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. *Medical Physics, 39*(7), 4093–4103.
Siegel, R. L., Miller, K. D., & Jemal, A. (2015). Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians, 65*(1), 5–29.
Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., & Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare Informatics Research, 16*(4), 253–259.
Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., et al. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences, 100*(18), 10393–10398.
SEER, (2017). Surveillance, Epidemiology, and End Results (SEER) program research data (1973–2014). In National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2017, based on the November 2016 submission. SEER. (www.seer.cancer.gov).
West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research, 162*(2), 532–551.
Wickramaratna, J., Holden, S., & Buxton, B. (2001). Performance degradation in boosting. In *Multiple classifier systems* (pp. 11–21). Springer.
Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In *Soft computing and industry* (pp. 25–42). Springer.
Zeng, T., & Liu, J. (2010). Mixture classification model based on clinical markers for breast cancer prognosis. *Artificial Intelligence in Medicine, 48*(2), 129–137.
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.
Zheng, B., Zhang, J., Yoon, S. W., Lam, S. S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications, 42*(20), 7110–7120.

Please cite this article as: H. Wang et al., A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2017.12.001