Computerized Medical Imaging and Graphics 38 (2014) 137–150

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

# Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD

Peng Cao[a,b,c,*], Jinzhu Yang[a], Wei Li[b], Dazhe Zhao[a,b], Osmar Zaiane[c]

[a] College of Information Science and Engineering, Northeastern University, Shenyang, China
[b] Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China
[c] Computing Science, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

*Article history:*
Received 13 May 2013
Received in revised form 19 October 2013
Accepted 2 December 2013

*Keywords:*
Lung nodule detection
False positive reduction
Imbalanced data learning
Ensemble classifier
Re-sampling
Random subspace method

ABSTRACT

Classification plays a critical role in false positive reduction (FPR) in lung nodule computer aided detection (CAD). The difficulty of FPR lies in the variation of the appearances of the nodules, and the imbalance distribution between the nodule and non-nodule class. Moreover, the presence of inherent complex structures in data distribution, such as within-class imbalance and high-dimensionality are other critical factors of decreasing classification performance. To solve these challenges, we proposed a hybrid probabilistic sampling combined with diverse random subspace ensemble. Experimental results demonstrate the effectiveness of the proposed method in terms of geometric mean (G-mean) and area under the ROC curve (AUC) compared with commonly used methods.



## 1. Introduction

Lung cancer is one of the main public health issues in developed countries, and early detection of pulmonary nodules is an important clinical indication for early-stage lung cancer diagnosis [1]. According to statistics from the American Cancer Society, lung cancer is the primary cause of cancer-related death in the United States [2]. Lung nodule refers to lung tissue abnormalities that are roughly spherical with round opacity and a diameter of up to 30 mm [3]. Currently, nodules are mainly detected by one or multiple expert radiologists inspecting CT images of lungs. Recent research, however, shows that inter-reader variability in the detection of nodules by expert radiologists may exist. In addition, since three-dimensional (3D) image processing and analysis techniques become applicable in thin-section CT, a thin-section CT scan includes hundreds of sections and requires considerable time and effort in image interpretation by radiologists. For more than a decade, significant effort has been focused on developing automated systems that detect/recognize suspicious lesions in thoracic CT imagery as well as other types of imagery. It is therefore an important task to develop computer aided detection (CAD) systems that can aid/enhance radiologist workflow and potentially reduce false negative findings. CAD is a scheme that automatically detects suspicious lesions (nodule, polyps and masses) in medical images of certain body part, and provides their locations to radiologists. Computer aided detection (CAD) has become one of the major research topics in medical imaging and diagnostic radiology, and has been applied to various medical imaging modalities including computed tomography (CT), magnetic resonance imaging, and ultrasound imaging [4–6]. Current CAD schemes for nodule characterization have achieved high performance levels and would be able to improve radiologists performance in the characterization of nodules in thin-section CT, whereas current schemes for nodule detection appear to report many false positives [4,7]. It is because detection algorithms have high sensitivity that some non-nodule structures (e.g., blood vessels) are labeled as nodules inevitably in the initial nodule identification step. Since the radiologists must examine each identified object, it is highly desirable to eliminate these false positives (FPs) as much as possible while retaining the true positives (TPs). Therefore, significant efforts are needed in order to improve the performance levels of current CAD schemes for nodule detection in thin-section CT [4].

The purpose of false-positive reduction is to remove these false positives (FPs) as much as possible while retaining a relatively high sensitivity. It is a binary classification between the nodule and non-nodule. In machine learning, the aim of classification is to learn a system capable of the prediction of the unknown output class of a previously unseen instance with a good generalization ability. The false-positive reduction step, or classification step, is a critical part

* Corresponding author at: Computing Science, University of Alberta, Edmonton, Alberta, Canada. Tel.: +1 5875579488.
*E-mail addresses:* pcao1@ualberta.ca, neusoftcp@gmail.com (P. Cao), yangjinzhu@neusoft.com (J. Yang), l-w@neusoft.com (W. Li), zhaodz@neusoft.com (D. Zhao), zaiane@cs.ualberta.ca (O. Zaiane).

0895-6111/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compmedimag.2013.12.003

in the lung nodule detection system [8–12]. In the last two decades, the use of data mining techniques has become widely accepted in the medical applications to support patient diagnosis more effectively. Mining medical images has been selected as one of the top 10 data mining case studies [13]. Classification is a common task in data mining. There are two significant problems in the classification of the potential nodules: one is the enormous variance in the volumes, shapes, and appearances of the suspicious nodule objects, it is difficult to construct a single classifier for describing and modeling the complex data; the other is that the two classes are skewed and have extremely unequal misclassification costs, which is a typical class imbalance problem [14,15]. The imbalanced data issue usually occurs in computer-aided detection systems since the healthy class is far better represented than the diseased class in the collected data [16,17], including other CAD, such as breast, colon [18–20]. Class imbalanced data has detrimental effects on the performance of conventional classifiers. Typically classifiers attempt to reduce global error rate without taking the data distribution into consideration. As a result, all instances are misclassified as negative for high classification accuracy. Recently, the class imbalance problem has been identified as one of the 10 main challenges of Data Mining [21]. To date, there is no systematic research about the class imbalance learning issue in the lung nodule detection.

Imbalance that exists between the instances of two classes is usually known as between-class imbalance. The actual cause for the poor performance of conventional classifiers on the minority class is not necessarily related to only the between-class imbalance. The existence of within-class imbalance is closely intertwined with the problem of small disjuncts, which has been shown to significantly decrease classification performance [22,23]. Within-class imbalance refers to the case where a class is formed of a number of sub-clusters with different sizes, concerns itself with the distribution of representative data for subconcepts within a class. The existence of sub-concepts also increases the complexity of the problem because the amount of instances among them is not usually balanced. It was verified to be more difficult to handle than datasets with only homogeneous concepts for each class. Within-class imbalanced data distribution may yield small disjuncts, which is the essential challenge in the within-class imbalanced data issue. A phenomenon sometimes referred to as the problem with small disjuncts and that these small disjuncts collectively contribute a significant portion of the total test errors [24]. Besides, high-dimensionality poses additional challenges when dealing with class-imbalanced prediction [25], it is often unavoidable to have data with high dimensionality and imbalanced class distribution; some specific examples include text classification and gene expression data analysis. The instances of minority class are prone to be sparse as dimensionality increases, resulting in amplifying the issue of imbalance data classification. The challenges above are also critical in the stage of false positive reduction of Lung nodule CAD. First, for nodule candidates data generated from the initial detection may contain several sub-concepts as both true and false nodule objects involve multiple different type or different characteristic, which results in the distribution of instances over each class concepts and may yield clusters with unequal sizes. Second, no single feature can discriminate the true and false nodules effectively, and it will produce an inadequate classifier if too few features are chosen. However, choosing too many features for characterizing potential nodule objects can induce high computation cost as well as the potential for overfitting. The complex data distribution of nodule candidate instances aggravates the recognition of nodule, since the sensitivity of traditional classifiers to class imbalance increases with the domain complexity and the degree of imbalance.

In order to solve the issues above, we propose a hybrid probabilistic sampling method combined with diverse random subspace ensemble algorithm (HPS-DRS). The hybrid probabilistic sampling (HPS) method adopts the combination of over-sampling and under-sampling, and incorporates probability function in its data distribution re-sampling mechanism. It generates more accurate instances to generalize the decision region for the nodule class, and removes the redundant instances for the non-nodule class without destroying the structure of the data. It can deal with the between-class imbalance and within-class imbalance issues simultaneously. In addition, to avoid the negative impact on the probability estimation due to the feature set and to improve the classification performance, we design a novel ensemble based on the random subspace method [26]. It not only injects more diversity into the ensemble via the learning algorithm, but also via the bias of the sampling algorithm, so as to acquire better classification performance and generalization capability. Furthermore, it can address the classification of high dimensional data and alleviate the negative influence due to the irrelevant and redundant features. To perform a rigorous validation with our system, we use multiple datasets including medical imaging datasets and UCI machine learning datasets. We empirically investigate and compare the proposed method with the state-of-the-art approaches in the class imbalance classification and the false positive reduction; experimental results show the unique feature of the proposed method for overcoming the challenges in the Lung nodule CAD and demonstrate the promising effectiveness of this method.

While several other CAD systems have been discussed in the literature, our approach has several novel aspects. We employ a fully automated algorithm for identifying and segmenting both lungs in the CT scans. Next, an ensemble-based re-sampling method is proposed, which can improve the performance of classification on the imbalanced data distribution data. To the best of our knowledge, the classification of nodule candidates in CAD system from the aspect of the characteristic of data distribution of nodule candidates, such as between-class or within-class imbalance, high dimensionality has not been previously reported.

The remainder of the paper is organized as follows: in Section 2 we review current state-of-the-art techniques for tackling the candidate nodule classification problem as well as the imbalanced data learning. In Section 3, we introduce the proposed method, hybrid probabilistic sampling combined with diverse random subspace ensemble. In Section 4 we present experimental results and draw our conclusion in Section 5.

## 2. Related work

In this paper, we focus on the potential nodule classification issue in the lung nodule CAD; thus we only review the existing lung nodule classification methods and the commonly used solutions for addressing the class imbalance problem.

### 2.1. The common methods for the nodule candidates classification

After the initial nodule identification step locates suspicious nodule candidates in CT images, the false-positive reduction step tries to classify the nodule candidates into nodule (positive class/minority class) and non-nodule (negative class/majority class) categories and, subsequently, to remove false positives by analyzing the features of nodule candidates. Classifiers are designed to generate models from sample data and the models are desired to best predict the future input data. Various classifier models have been applied for reducing the false positive nodules. One of the most frequently employed and simplest classifier is the rule-based classifier [27]; however it is hard to determine the selection of cut-off threshold to classify abnormal and normal manually. Since linear discriminant analysis (LDA) offers simplicity in computation and effectiveness in classification, it is commonly used to discriminate

the potential nodule [11,28]. More sophisticated classifiers such as neural network (NN) and support vector machine (SVM) are often employed in nodule recognition tasks [8,10,29–31], which have the ability to learn complex input–output relationships automatically, and have low dependence on domain specific knowledge. An important new trend is the appearance of ensemble learners which combine the decisions of multiple classifiers to form an integrated output, so as to enhance the generalization ability of a single model. Suzuki et al. have proposed a pixel-based massive training artificial neural network (MTANN) for distinction between nodule and FPs [10]. Lee et al. developed a random forest ensemble classification to improve the nodule classification performance [32]. Dolejsi et al. designed a Asymmetric Adaboost ensemble to reduce the number of FPs [33].

There is an important problem in the classification of potential nodule data. The dataset is typically imbalanced, and the costs of misclassification are different. Class imbalanced data has detrimental effects on the performance of conventional classifiers, resulting in lowering the performance of discrimination in the candidate nodule. However, in nodule classification, the problem has attracted less attention. Only a few publications address this problem. The authors in [8] use Tomek links to remove borderline false nodule cases in order to achieve 100% sensitivity. Campadelli et al. prove that cost-sensitive SVM (CS-SVM) trained with imbalanced data sets achieves promising results in terms of sensitivity and specificity, by means of adjusting the misclassification cost of false positives versus false negatives [31]. Dolejsi et al. use asymmetric Adaboost learning to improve the sensitivity by setting different weights for two classes [33].

### 2.2. The common methods for the class imbalance problem

Not only in the medical lesion detection domain, many real-world applications, such as spam filtering, text classification and fraud detection in business transactions, have problems when learning from imbalanced data sets. In recent years, the imbalanced learning problem has drawn a significant amount of interest from academia, industry, and government funding agencies. Much work has been done in addressing the class imbalance problem. These methods can be grouped in two categories: the data perspective and the algorithm perspective [15]. The methods with the data perspective re-balance the class distribution by re-sampling the data space, either over-sampling instances of the minority class or under-sampling instances of the majority class. The re-sampling techniques try to balance out the dataset either randomly or deterministically. A widely used over-sampling technique is called SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic samples between each positive sample and one of its neighbors [34]. SMOTE is effective to increase the significance of the positive class in the decision region. There exist many methods based on the SMOTE for generating more appropriate instances [35,36]. The methods with the algorithm perspective adapt existing common classifier learning algorithms to bias toward the small class, such as one-class learning and cost sensitive learning. Cost-sensitive learning is one of the most important topics in machine learning and data mining, and has attracted high attention in recent years. It takes misclassification costs into account during the model construction, and does not modify the imbalanced data distribution directly [37–39].

As we have stated, in recent years, ensemble of classifiers have arisen as a possible solution to the class imbalance problem attracting great interest among researcher because of their flexible characteristics [40]. Ensembles are designed to increase the accuracy of a single classifier by training several different classifiers and combining their decisions to output a single class label. Not only multiple classifiers could have better answer than a single one, but also the ensemble framework provides diversity for avoiding the overfitting of some algorithms. Bagging and Boosting are two of the most popular techniques; the Easyensemble method is developed based on the Bagging classification [41]. SMOTEBoost [36] is designed to alter the imbalanced distribution based on Boosting. Data generation techniques are involved to emphasize the minority class examples at each iteration of Boosting. The ensemble combined with cost sensitive learning can enhance the performance due to the diversity of the ensemble, such as AdaCost [42]. and MetaCost [43].

## 3. HRS-DRS method

In this section, we begin by describing the hybrid probabilistic sampling. We then describe how to incorporate this method into the diverse random subspace ensemble to create HPS-DRS.

### 3.1. Hybrid probabilistic sampling

Gaussian mixture models (GMM) are generative probabilistic models of several Gaussian distributions for density estimation in machine learning applications. A Gaussian mixture can be constructed to acceptably approximate any given density. Therefore, we assume the distribution of two classes follows the Gaussian mixture model with unknown parameters. The parametric probability density function of GMM is defined as a weighted sum of Gaussians. The finite Gaussian mixture model with $k$ components may be written as:

$$p(y|\mu_1, \ldots, \mu_k; \sigma_1, \ldots, \sigma_k; \pi_1, \ldots, \pi_k) = \sum_{j=1}^{k} \pi_j N(\mu_j, \sigma_j) \quad (1)$$

and

$$0 \le \pi_j \le 1, \quad \sum_{j=1}^{k} \pi_j = 1 \quad (2)$$

where $\mu_j$ are the means, $\sigma_j$ are covariance matrixes, $\pi_j$ are the mixing proportions, and $N(\mu_j, \sigma_j)$ is a Gaussian with specified mean and variance.

We need to estimate the parameters of GMM with the existing instances of both the classes. The standard method used to fit finite mixture models to observe data is the expectation-maximization (EM) algorithm, which converges to a maximum likelihood estimate of the mixture parameters. However, the drawbacks are that it is sensitive to initialization and it requires the number of components to be set by users. Since the FJ algorithm [44] tries to overcome the major weaknesses of the basic EM algorithm particularly vis-à-vis the initialization, and can automatically select the number of component, we use it here to estimate the parameters of GMM.

Each instance $x_i$ will then be assigned to the cluster $k$ where it has the largest posterior probability $p(k|x_i)$. When calculating the probability of each instance on each component, the probabilities for the numeric attributes is obtained by a Gaussian density function, and for the nominal attributes, the probabilities of occurrence of each distinct value are determined using Laplace estimates. At the same time, we obtain the parameters of each Gaussian component. For different clusters, the re-sampling rates are different; within the cluster, the probabilities of each instance to be chose for re-sampled are different.

We use the over-sampling combined with under-sampling to balance the class size. The sizes of the two classes are $M_{neg}$ and $M_{pos}$. The gap $G$ between two uneven classes is: $G = M_{neg} - M_{pos}$. Thus, the amount of instances in the positive class for over-sampling is: $N_{pos} = G \times \alpha$, and the amount of instances in the negative class for under-sampling is: $N_{neg} = G \times (1 - \alpha)$. To adjust the within class imbalance, we need to balance cluster sizes in each class. For the positive class, the number of instances to be over-sampled is

inversely proportional to the size of the cluster; for the negative class, the numbers of instances to be under-sampled are proportional to the size of the cluster. For example, there are three clusters of size 20, 15 and 10 in the negative class, and two clusters of size 10 and 5 in the positive class. If $\alpha$ is set to 50%, the gap $G$ is 30, $N_{pos} = N_{neg} = 15$. The sizes of the three clusters in the negative class become 13, 10 and 7 after under-sampling, while both the sizes of the two clusters in the positive class become 15 after over-sampling. This reduces the within class imbalance, and in this case equalizes the class sizes.

Furthermore, we use the probabilities of each instance to conduct the re-sampling with maintaining the data structure, in order to address the two type imbalance issues. In the clusters of the negative class, the instances with higher probability are dense, they are frequent in the subclass, and hence they have higher chance to be under-sampled. We choose the instances to be under-sampled according to the Gaussian distribution. In the clusters of the positive class, the new instances are produced according to the probability function of Gaussian distribution, resulting in finding more potentially interesting regions. The main steps in under-sampling for the clusters of the negative class and over-sampling for the clusters of the positive class according to the distribution probability are the following:

#### 3.1.1. Over-sampling phase

**Step 1**: In the over-sampling for the positive class, the smaller the size of cluster within the class, the more instances are over-sampled, so as to avoid the small disjuncts. For the $i$th cluster, the amount of synthetic instances needed to be generated is:

$$N_{pos}^{i} = \left( \frac{1/size_{pos}^{i}}{\sum_{j=1}^{S_{pos}} (1/size_{pos}^{j})} \right) \times N_{pos} \tag{3}$$

where $size_{pos}^{i}$ is the size of $i$th cluster, $S_{pos}$ is the number of clusters in the positive class.

**Step 2**: In the $i$th cluster, $N_{pos}^{i}$ instances are generated with the parameters from the current Gaussian distribution. The new instances are generated according to the probability function of the Gaussian distribution with parameters learned from the available data. Firstly, the probability from the Gaussian distribution of each instance is calculated and normalized:

$$\hat{p_k} = \frac{p_k}{\sum_{j=1}^{size_{pos}^{i}} p_j} \tag{4}$$

Then, the amount of new instances for each instance $x_k$ is obtained according to:

$$n_k = N_{pos}^{i} \times \hat{p_k} \tag{5}$$

For ensuring that synthetic instances created via this method always lay in the region near $x_k$, the $n_k$ instances are generated in its $K$ nearest neighbors region. It can extend more potential regions rather than being limited along the line between the positive example and its selected nearest neighbors. In addition, this guarantees the creation of positive samples in the cluster, and avoids any incorrect synthetic instance generation.

#### 3.1.2. Under-sampling phase

**Step 1**: In the under-sampling for the negative class, we calculate the amount of instances to be under-sampled for each cluster. The number of instances to be under-sampled are proportional to the size of clusters. For the $i$th cluster, the amount of instances needed to be removed is:

$$N_{neg}^{i} = \left( \frac{size_{neg}^{i}}{\sum_{j=1}^{S_{neg}} size_{neg}^{j}} \right) \times N_{neg} \tag{6}$$

where $size_{neg}^{i}$ is the size of $i$th cluster, $S_{neg}$ is the number of clusters in the negative class.

**Step 2**: In each component Gaussian distribution, the center region is denser than the border region. These instances from the center are more possible to be redundant, and so are better candidates to be under-sampled. We need to choose the instances to be ignored or removed located on the center of the distribution more than the border. The probabilities to be chosen for under-sampling are proportional to the normalized probability $\hat{p}$ of the Gaussian distribution for each instance in a cluster.

Before applying GMM, to avoid the effect of noise instances, we filter out the noise by checking the labels of nearest neighbors. We remove any noisy example which violates the rule that the class label of each instance is consistent with the one of at least three of its five nearest neighbors.

The procedure described above is the main scheme of HPS. This general idea of HPS and the difference of this idea to SMOTE are visualized in Fig. 1. The (a) is the original skewed data distribution. We can see the positive class has two subclasses with within-class imbalance and an outlier instance. These factors may decrease effectiveness of the learning and over-sampling. The (b) is the result of the SMOTE. The procedure of SMOTE conducts the linear interpolation between nearest neighbor instances, resulting in generating many wrong positive instances under the complex distribution. We see that, some wrong positive samples are interpolated into the region of the negative class since noise and class dispersion exist. Hence, it is not sufficient to manipulate the class size without considering the local distribution. The (c) and (d) show the strategy of our HPS. The clustering result of GMM is shown in (c). (d) is the final result of the HPS. We can see that HPS is able to broaden the decision regions and the concept of positive class from a global perspective to a perspective that encompasses local information in order to deal with within-class imbalance.

### 3.2. Integration of HPS and diverse random subspace, HPS-DRS

The redundancies and noise in the feature set hinder the re-sampling techniques to achieve their goals. Moreover, the quality of probability estimation and classification will largely depend on the feature set. The irrelevant or redundant features can lead to a decrease in performance on the re-sampling and prediction.

An important trend in machine learning is the appearance of ensemble learning which combines the decisions of multiple weak classifiers to form an integrated output, so as to provide a diversity for avoiding the overfitting for some algorithms. Moreover, ensemble learning is also a good solution for solving the class imbalance problem, as it is incorporated with re-sampling technique to acquire better classification performance and generalization capability. Ho showed that the random subspace method is able to improve the generalization error [26]. In the random subspace ensemble, the individual classifier is built by randomly projecting the original data into subspaces and training a proper base learner on these subspaces to capture possible patterns that are informative on classification. The majority voting scheme is utilized when combining each specific classifier's prediction.

Under the current standard random subspace scheme, there are three disadvantages requiring improvement: (1) it only picks the feature subset for the original feature set randomly without considering the diversity of instances. Projecting the feature space on a given subspace could produce or enhance noisy instances and even contradicting instances that would lead to poor performance. This is the case when values of attributes in the selected subspace are outliers. (2) Since the features for a classifier are selected independently from the feature subspaces of other classifiers in the ensemble, the standard RS scheme has random characteristics

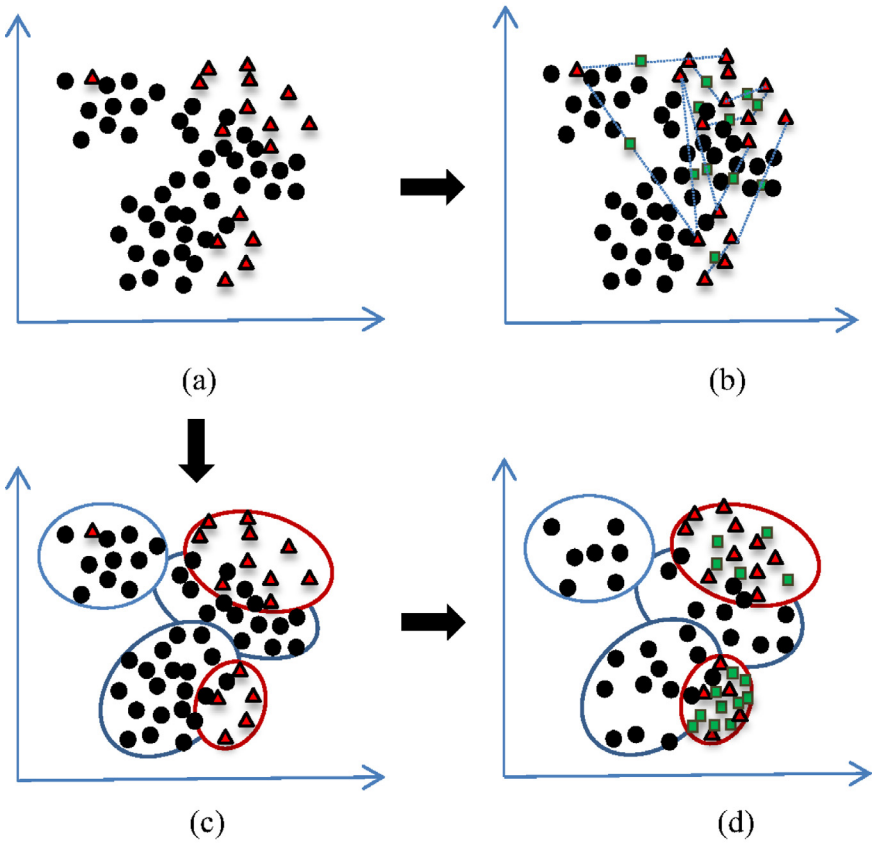



**Fig. 1.** Comparison of different synthetic data generation mechanisms. (The black circles and the red triangles represent the negative and positive classes, respectively. The green squares are the new instances generated by over-sampling). (a) Original imbalanced data distribution. (b) Data distribution after SMOTE. (c) The result of Gaussian mixture clustering. (d) Data distribution after HPS. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

through the selection of feature subsets. However, there are still strong overlaps of the instances with feature selected when constructing individual classifiers on different subspaces, as there is no formulation to guarantee small or reduced overlap. (3) Because some subspaces may contain noisy features and individual classifier developed from these subspaces are not informative, it is not correct treating each classifier as if it contributed equally to the group's performance; there is a lack of attention to appropriate weight assignments to individual classifiers according to their respective performance based on the different subspace.

Therefore, we propose an improvement of RS, called DRS (diverse random subspace) for addressing these disadvantages. Firstly, we extend the common random subspace by integrating bootstrapping samples in order to obtain the diversity with respect to instances and features. In the bootstrapping method, different training subsets are generated with uniform random selection with replacement. Secondly, it cannot ensure the diversity of each subset since the instances and the features are chosen randomly without considering previously selected subspaces for other classifiers. Therefore, to improve diversity between each subset, we use a formulation to make sure each subset is diverse. We introduce a concept of *overlappingrate*:

$$overlappingrate = \frac{subset_i \cap subset_j}{N_{fea} \times N_{ins}} \tag{7}$$

where the $subset_i$ and $subset_j$ are two subsets within certain subspaces, $N_{fea}$ and $N_{ins}$ are the feature size and instance size of each subset; e.g., in Fig. 2, the overlapping rate is 16%.

We then introduce a threshold $T_{over}$ to control the intersection between each subset. The overlapping rate of all the subsets should be smaller than the threshold $T_{over}$.

The *GenerateDiverseSets* described in Algorithm 1 generates a diverse set *DiverseSet*, by iteratively projecting bootstrap sample $D_k$ into the specific random subspace $RS(D_k)$. The function $isDiverse(RS(D_k), DiverseSet, T_{over})$ examines if the new projection $RS(D_k)$ is diverse enough from the previously collected projections in *DiverseSet* based on the overlapping region threshold $T_{over}$. The generation of projections stops when there is stagnation *sr*, after enough trials, no new projection is diverse enough from the collected subsets. It enforces the diversity or independence by minimizing the overlapping region among the subset with subspace used previously.

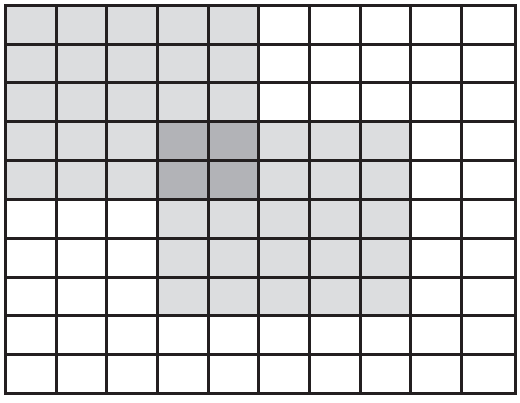


**Fig. 2.** The overlapping rate between two subsets.

**Algorithm 1.** [GenerateDiverseSets]

**Require:**
Training dataset, $D_{train}$
Ratio of bootstrap samples, $R_s$
Ratio of feature subspace, $R_f$
Overlapping region threshold, $T_{over}$
Stagnation rate, $sr = 100$

**Ensure:**
Diverse dataSets, *DiverseSets*

```
change = 0 ; DiverseSet = { }
while change < sr do
    A bootstrap sample D_k selected with replacement from D_train with R_s
    Select an random subspace with R_f from D_k
    if isDiverse(subspace(D_k), DiverseSet, T_over) == true then
        DiverseSet . add(subspace(D_k)) ; change = 0 ;
    else
        change = change + 1 ;
    end if
end while
```

Thirdly, we employ a weighted average while combining classifiers according to the performance of each component. In the diverse subsets, some of the selected subspaces may have better performance on the imbalanced dataset; others lack the ability to properly discriminate between the different classes. We utilized the out-of-bag (OOB) samples in determining different classifier's voting power, and then each base classifier is weighted when combined to create the final decision function. The goal is to assign weights that reflect the relative contribution of each classifier in improving the overall performance of the ensemble. It is known that the use of overall accuracy is not an appropriate evaluation measure for imbalanced data, therefore the metric for representing the performance of each classifier is chosen by G-mean. The G-mean is the geometric mean of accuracies measured separately on each class, which is commonly utilized when performance of both classes is concerned and expected to be high simultaneously.

$$G\text{-}mean = \sqrt{sensitivity} \times specificity \tag{8}$$

$$sensitivity = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad specificity = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{9}$$

where TP denotes the number of true positives, FP denotes the number of false positives, FN denotes the number of false negatives, and TN the number of true negatives.

The HPS-DRS algorithm is described in Algorithm 2.

**Algorithm 2.** [HPS-DRS]

**Require:**
Training Dataset $D_{train}$, Test Dataset $D_{test}$, Ratio of bootstrap samples $R_s$, Ratio of feature subspace $R_f$, Overlapping region threshold $T_{over}$, Hybrid sampling ratio parameter $\alpha$

**Training:**

```
Ensemble = NULL
DiverseSets=GenerateDiverseSets(D_train,R_s,R_f,T_over)
for each subset D_k in DiverseSets do
    Apply HPS on the subset D_k, and generate a new balanced set BD_s^k with α
    Construct a classifier C_k on the BD_k
    Evaluate C_k on the OOB(D_k) and obtain the value of G-mean, GM_k
    C_k . Subspace = Subspace(D_k); C_k . GM = GM_k
end for
Ensemble = Ensemble ∪ C_k
Calculate and normalize the weights of each classifier in Ensemble according to its GM
```

**Testing:**

```
Calculate output from each classifier of Ensemble with D_test
Generate the final output by aggregating all the outputs with weighted voting
```

To reduce the learning time of HPS-DRS, the procedure of sampling and learning in each subset $D_k$ can be carried out in parallel before aggregating. Moreover, each classifier is trained in the reduced subset with fewer instances and features. Therefore the computational time is acceptable.

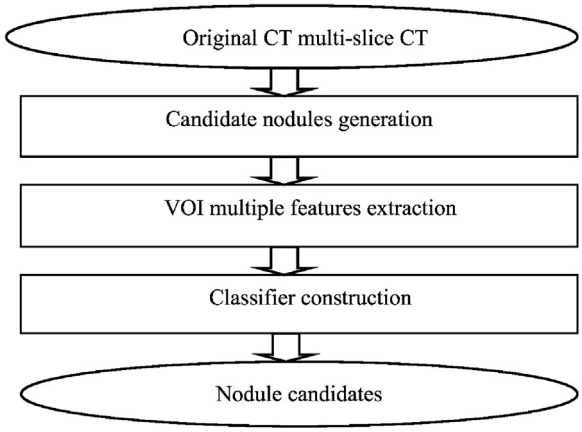



**Fig. 3.** The scheme of the nodule detection.

Since data often exhibits characteristics at a local rather than global level, DRS can find more valuable local data properties so as to improve the quality of sampling. Moreover, the different imbalanced data distribution in each random subset makes the ensemble classifier robust to the evolving testing distribution. Furthermore, DRS can alleviate the effect of class overlapping on the imbalanced data distribution [45], since the two classes may be separable in some reduced subspace.

## 4. Experimental study

### 4.1. CAD system

Unlike other publications in which candidate nodules were selected manually for classification evaluation, we obtained the appropriate candidate nodule samples objectively using a candidate nodule detection algorithm. Our CAD scheme also contains lung segmentation, candidate nodule detection and VOI segmentation. The sequence of steps for our scheme is shown in Fig. 3.

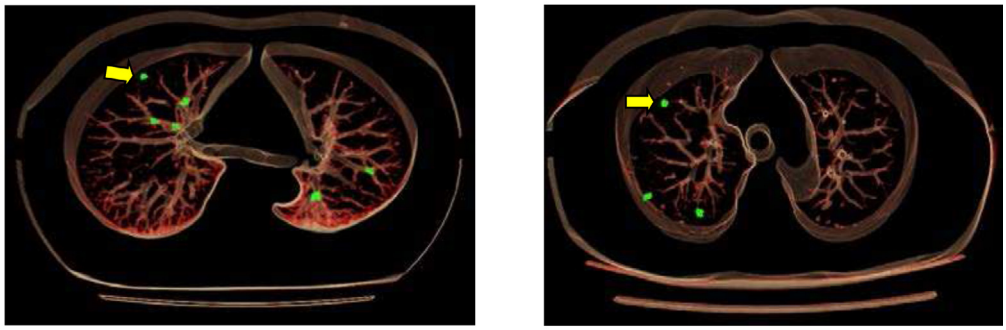
### 4.2. Initial nodule detection

Our database consists of 165 thin section CT scans with 192 solid nodules, obtained from the Guangzhou hospital and the Shenyang hospital in China. These databases included nodules of different sizes (3–30 mm). The nodule locations of these scans are marked consensually by three experienced expert radiologists.

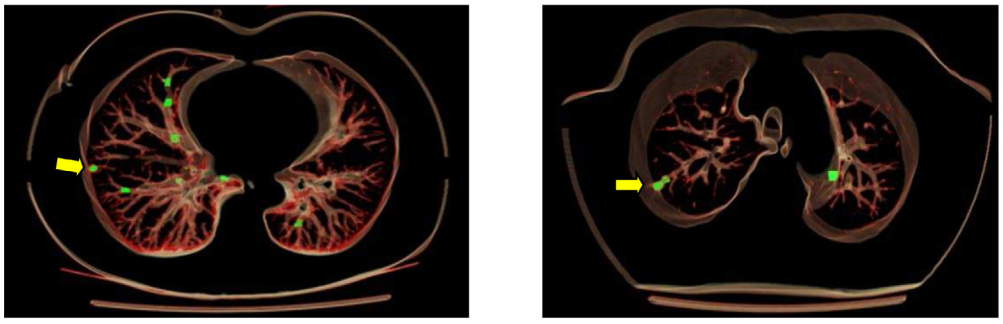
In the detection phase, we use the dot enhancement filter proposed by Li [46], which is aimed to simultaneously enhance objects of a specific structure (e.g. dot-like nodules) and suppress objects of other structures (e.g. line-like vessels) in multi-scale. If the nodule diameters are in the range $[d_{min}, d_{max}]$, the scales to be considered for the Gaussian smoothing will be in the range $[d_{min}/4, d_{max}/4]$. The amount of scales is set to 5. After the 3D selective enhancement Hessian filter is applied on the original image, we use a thresholding technique and a 3D connected-component labeling technique to separate nodule candidates and identify all of the isolated objects [47]. The value of the threshold in the thresholding technique is set to 40, and objects with an effective diameter smaller than 2.5 mm were considered to be non-nodules and were eliminated [47].

For each nodule candidate object, we developed a 3D constrained region-growing technique to segment it accurately in the original CT images. Fig. 4 shows some example result images of candidate VOI detection. The detected true nodule has different location and connection with the surrounding pulmonary structures.

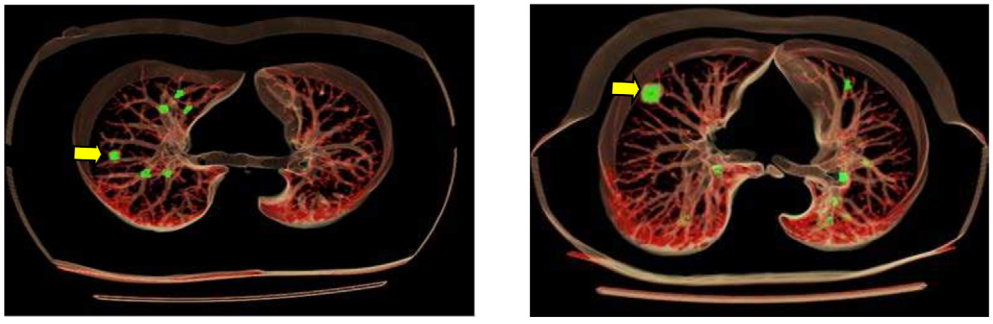
The total number of the potential nodule VOI segmentated by our methods is 1045. For the purpose of training the classifier with the nodule candidate dataset, we have adopted a simple candidate



(a) well-circumscribed nodule



(b) juxta-pleural nodule



(c) juxta-vascular nodule

**Fig. 4.** Initial detection result of candidate nodules. TP indicated by arrow, other spots are FP.

labeling rule for the detected nodule candidates to be classified as nodules or non-nodules based on annotation provided by the chest radiologists. During the evaluation, each detected nodule candidate was determined to be a nodule if its distance to target nodule in the database was smaller than 5 mm. We denote this classification as a hit. If a hit on a detected nodule candidate was produced, it was counted as a nodule (TP); otherwise, it was considered a non-nodule (FP). As a result, the detection algorithm identifies 171 true nodules as positive class and 874 non-nodules as negative class from the total CT scans according to the ground truth; the class imbalance ratio is 5. After obtaining the candidate nodule instances, we calculate the features for each of the nodule candidates from intensity, shape and gradient distribution aspects. Then, we need to reduce the false positive nodule instances with our proposed classification model.

### 4.3. Feature extraction

Feature extraction plays an important role in classification of suspicious nodule [11,12]. However, there is not a single outstanding feature that can discriminate the nodule from non-nodule completely. This is due to the fact that the nodules vary enormously in volume, shape, and appearance, and the sources of false positives are different. The majority of false positives are mainly caused by blood vessels and other normal anatomic structures. Some of the false positives can be easily distinguished from true nodule, however, a large portion of them are difficult to distinguish. Therefore, for getting a high classification accuracy in candidate nodule classification, we should extract more features from many aspects, such as intensity, shape and gradient distribution. Our feature extraction process generated 43 image features from volume of interest (VOI) for each potential nodule object. Using these features, we construct the input space for our classifiers. This section gives a brief introduction to the features we have collected for analysis and selection.

#### 4.3.1. Intensity feature

The nodules often have higher gray values than parts of vessels misidentified as nodules; and the intensity distribution of nodules CT appearance can be approximated by a Gaussian function. Therefore, we use the some statistical gray feature to describe the global gray distribution within the candidate objects [8]. The gray value within the objects was characterized by use of seven statistics (mean, variance, max, min, skew, kurtosis, entropy).

Meanwhile, for capturing the volume intensity distribution along the radial direction of the three dimension nodule candidates, we extract the feature of radial intensity distribution (RIS) [48]. The vector of radial intensity distribution feature for nodule object should be decreasing, while for non-nodule it changes irregularly. It is computed as follows:

$$RIS(r_{i+1}, r_i) = \sum_i I(x_i, y_i, z_i) \tag{10}$$

where $I(x_i, y_i, z_i)$ denotes the intensity value in the position of $(x_i, y_i, z_i)$, and $(x_i, y_i, z_i) \in \{(x_i, y_i, z_i) | r_i < \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2 + (z_i - z_c)^2}\}$, $(x_c, y_c, z_c)$ is the coordinate of the center.

#### 4.3.2. Shape feature

Based on the fact that an isolated nodule or a nodule attached to a blood vessel is generally either depicted as a sphere or has some spherical elements, while a blood vessel is usually oblong. We attempt to distinguish true nodules from false ones by calculating the volumetric shape index (SI) and curvedness (CV) [49–51], as well as other explicit shape features to characterize the 3D shape [47].

Shape index is a measure of the shape, which represents the local shape feature at each voxel while being less sensitive to the image intensity. Every distinct shape, except for the plane, corresponds to a unique SI. The volumetric shape index (SI) at each voxel $p$ can be defined as:

$$SI(p) = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)} \sqrt{\frac{k_1^2 + k_2^2}{2}} \tag{11}$$

$$CV(p) = \sqrt{\frac{k_1^2(p) + k_2^2(p)}{2}} \tag{12}$$

The SI and CV features are calculated based on every pixel, and they are further characterized by seven statistical operations (mean, variance, max, min, skew, kurtosis, entropy). Some other explicit shape features were also extracted from candidate object, including the volume, the surface area and compactness. Features 13–29 describe the intensity distribution from the VOI in Table 1.

#### 4.3.3. Gradient distribution feature

The true nodules have a high concentration because they grow from the center to surround, thus nodules have high concentration of gradient vector. We utilize the concentration feature to characterize the degree of convergence at voxel $p$ based on the concept of gradient concentration (GC) [49,52]. In short, the GC feature at a point $p$ characterizes the overall direction of the gradient vectors around. It is defined by:

$$GC(p) = \frac{1}{D} \sum_{i=1}^{D} e_i^{\max}(p) \tag{13}$$

$$e_i^{\max}(p) = \max_{R_{\min} \le n \le R_{\max}} \left\{ \frac{1}{n+1} \sum_{l=1}^{N} \cos \theta_{il}(p) \right\} \tag{14}$$

where $D$ is the number of the symmetrically direction vectors $d^l$ originating from $p$. The angle is calculated between dl and $g_i^l$, where $g_i^l$ is the gradient vector located at distance $l$ from $p$ in direction $d^l$. We also calculate the gradient field strength at each pixel. The concentration and strength features based on every pixel are further characterized by seven statistical operations. Features 30–43 describe the gradient distribution from the VOI in Table 1.

### 4.4. Evaluating the effectiveness of HPS-DRS

We made a vertical comparison and a horizontal comparison separately. In the vertical comparison, we tested the HPS-DRS by comparing our hybrid approach to each individual technique as well as original RS and Bagging with and without the re-sampling. In the horizontal comparison, we compared the HPS-DRS with the state-of-the-art methods in the classification of imbalanced data as well as the current methods for false positive reduction.

In this experiment, we evaluate the effectiveness of our proposed HPS-DRS algorithm. Since the DRS framework utilizes the idea of RSM and Bagging, we conduct the comparison between DRS, original random subspace, Bagging as well as the single methods with or without HPS. We chose neural network (NN) as our base classifier. In the setting of the neural network classifier, the number of input neurons is equal to the number of features in the given subspace, and the number of neurons in the hidden layer is set to be 10. The sigmoid function is used as the activation function, and the inner training epochs is set to be 200 with a learning rate of 0.1.

We use metrics such as sensitivity, specificity and G-mean to evaluate the performance of the learning algorithm on imbalanced nodule candidate data. To make our comparisons more convincing, we further use the AUC (area under the ROC curve) as the performance evaluation [30,53] which is a commonly used measurement in medical CAD systems. The AUC measures the performance of ranking a randomly chosen positive example higher than a randomly chosen negative example. In this case, it represents the performance of ranking an instance from the positive class higher than instances in the negative class. The results show that three ensemble approaches increase performance of the single model by using multiple different and complementary representations. We also use AUC as the measure metric of the weighted voting in the aggregation of DRS ensemble. The ensemble size of all the ensemble methods is set to 50. In the parameters setting of HPS, the $\alpha$ is set to 70% for avoiding loss of reducing too many instances, and the parameters $K$ set to 5. In the parameters setting of DRS, $R_s$ is 0.7, $R_f$ is 0.5. The best value of $T_{over}$ can be obtained from the training data, then the ensemble size can be determined adaptively. In this experiment, it is set to 0.4 empirically. It is a good trade-off value between the diversity and the sufficient ensemble size according to experiments. Although it is not necessarily the best, it can guarantee the diversity among each subset. All the experiments are carried out by means of a 10-fold cross-validation. That is, the dataset was split into 10-fold, each one containing 10% of the patterns of the dataset. For each fold, the algorithm is trained with the examples contained in the remaining folds and then tested with the current fold. The results are shown in Table 2.

From Table 2, we can observe that the performance of neural network (NN) is affected by the presence of class imbalance. It is also apparent that the proposed HPS can improve the performance of the original NN on the imbalanced nodule candidate data regardless of the single model or ensemble, which can show that HPS has the ability of reducing the bias inherent in the learning procedure due to the class imbalance. However, the classification performance based on only HPS did not achieve the best, since the over-sampling performed on the whole space cannot get the optimal re-sampling quality due to the fact that redundant or irrelevant features exist.

The experimental results also demonstrated that the Diverse Random Subspace (DRS) ensemble framework, especially designed for imbalanced problems, obtained a more generalized performance for high-dimensional and imbalanced potential nodule data. The mechanism of aggregating each component with weighted voting power using G-mean is beneficial for the imbalanced data, because it treat each classifier unequally and assign the appropriate weight to each classifier based on each performance on the imbalanced data. Moreover, the distribution of the class overlapping

**Table 1**
Feature set for candidate nodule classification.

| ID | Category | Feature | Description |
|---|---|---|---|
| F1–7 | Intensity | Intensity statistical feature | The gray value within the objects was characterized by use of seven statistics (mean, variance, max, min, skew, kurt, entropy) |
| F8–12 | | Sub-volume distribution feature | The average intensity within each sub-volume along the radial directions |
| F13–19 | Shape | SI statistical feature | The volumetric shape index (SI) representing the local shape feature at each voxel was characterized by use of seven statistics |
| F20–26 | | CV statistical feature | The volumetric curvedness (CV), which quantifies how highly curved a surface is, was characterized by use of seven statistics |
| F27–29 | | Geometric feature | Volume, surface area and compactness |
| F30–36 | Gradient | Concentration statistical feature | The concentration characterizing the degree of convergence of the gradient vectors at each voxel, was characterized by use of seven statistics |
| F37–43 | | Gradient strength statistical feature | The gradient strength of the gradient vectors at each voxel, was characterized by use of seven statistics |

differs in all data subsets, and this causes the diversity to be boosted, so as to avoid overfitting and increase the generalization ability by constructing a number of assorted and complementary. Although single DRS can obtain the best specificity, the improvement of sensitivity is limited as no new positive instances are introduced.

HPS-DRS can combine the advantage of two techniques from two aspects of adding useful information and enhancing generalization performance, so obtain an improvement in prediction performance on the positive class (sensitivity) and overall improved G-mean as well as AUC. The hybridization of data preprocessing techniques with diverse ensemble framework achieves better performance with less classifiers, which demonstrates the appropriateness of the hybrid use instead of applying a unique preprocessing re-sampling step and training a single neural network classifier. The reason is that: first, the ensemble framework of DRS can inject more diversity into the bias of sampling and learning algorithm, so as to help HPS acquire better over-sampling performance and achieve a better generalization ability. This reduces the possibility of re-sampling distorting the data distribution so that the classifier is prevented from over-fitting. Second, the hybrid probabilistic sampling method can gain more when working in a reduced feature space and can find local potential interesting pattern, since the complex data often exhibits characteristics and properties at a local, rather than global level.

The other common ensemble methods, such as Bagging and RSM, perform worse than DRS regardless of using re-sampling or not, since Bagging or RSM works on the entire feature space or on the entire instance space, which cannot find potential local data property and cannot avoid the noise of instance or feature. We posit that by first randomly selecting multiple reduced data subset with reduced subspaces, sampling along that diverse subsets, and then learning classifier will induce a higher diversity, which is a necessary condition for improved performance of classifier. Additionally, the classic Bagging and RSM are both accuracy-oriented designs, hence they do not solve the problem that underlay in the base classifier by themselves when directly applied to imbalanced data sets, whereas the mechanism of aggregating each component with weighted voting power using G-mean is effective on the imbalanced data.

In addition to the Gaussian mixture distribution, we tried other distributions for continuous variable to model the nodule candidate data, including single Gaussian distribution, Laplace distribution, Exponential distribution, and Weibull distribution. The parameters of the assumed distribution of both classes are learned individually based on the training data firstly, then for the positive class, synthetic data points are introduced into the original data set by generating from the learned probability distribution; for the negative class, the reduandant instances are under-sampled based on the probabilities of each instance, so as to keep the important information as well as the structure of data distribution. From Table 3, the re-sampling based on other distribution cannot achieve acceptable results, which demonstrates that our proposed re-sampling algorithm is dependent upon the chosen probability distribution, and is effective only if the chosen probability distributions can model the data appropriately. The comparative results empirically show that the feature extracted from nodule candidate data can be well fitted by Gaussian mixture distribution.

By the vertical comparison, our methods have been shown to improve on the performance of reducing false positives while maintaining the high recognition of true nodules i.e. our methods do not jeopardize the high sensitivity of the initial detection.

### 4.5. The effect of re-sampling ratio on the performance

In HPS, the optimal re-sampling ratio $\alpha$ may be unknown, and the parameter plays a vital role for the performance of hybrid re-sampling on the imbalanced data learning. In the previous experiments, $\alpha$ is set to 70% empirically. The experiment shows the performance by tuning the re-sampling ratio. The range of $\alpha$ is [0, 100%]; the step is set to 10%. With each value of $\alpha$, we conduct 10-fold cross validation to obtain an averaged G-mean and AUC results. From Fig. 5, we can see the changes of G-mean and AUC when varying the value of $\alpha$. When $\alpha$ is 0, only under-sampling for negative class is carried out and no new instances are generated.

**Table 2**
The results of three ensemble classifiers and single classifier.

| | Method | Sensitivity (%) | Specificity (%) | G-mean (%) | AUC (%) |
|---|---|---|---|---|---|
| Single | NN | 71.38 ± 6.93 | 88.75 ± 5.13 | 79.53 ± 6.23 | 81.24 ± 3.46 |
| | HPS + NN | 78.47 ± 4.72 | 89.17 ± 5.29 | 83.58 ± 5.35 | 84.32 ± 2.93 |
| RSM | NN | 75.25 ± 3.54 | 91.89 ± 3.81 | 83.09 ± 4.17 | 83.65 ± 3.37 |
| | HPS + NN | 79.64 ± 3.27 | 90.93 ± 2.97 | 85.06 ± 3.03 | 88.70 ± 4.42 |
| Bagging | NN | 73.43 ± 1.55 | 91.15 ± 1.89 | 81.77 ± 2.25 | 83.14 ± 2.01 |
| | HPS + NN | 80.50 ± 3.61 | 88.67 ± 3.28 | 84.45 ± 3.19 | 86.47 ± 2.86 |
| DRS | NN | 76.97 ± 4.36 | **92.29 ± 5.25** | 84.20 ± 5.11 | 85.97 ± 5.07 |
| | HPS + NN | **84.23 ± 3.14** | 91.50 ± 4.49 | **87.77 ± 4.62** | **91.65 ± 3.25** |

**Table 3**
The results of re-sampling based on different distribution.

| Distribution | Method | Sensitivity (%) | Specificity (%) | G-mean (%) | AUC (%) |
|---|---|---|---|---|---|
| Single Gaussian distribution | HPS | 70.84 ± 6.65 | 82.75 ± 6.24 | 76.56 ± 6.77 | 79.92 ± 5.32 |
| | HPS-DRS | 71.55 ± 6.48 | 83.44 ± 5.57 | 77.27 ± 5.95 | 79.33 ± 4.74 |
| Laplace distribution | HPS | 72.65 ± 3.22 | 85.89 ± 2.70 | 78.99 ± 3.09 | 81.23 ± 2.76 |
| | HPS-DRS | 76.37 ± 4.11 | 86.78 ± 3.35 | 81.41 ± 3.62 | 83.19 ± 2.77 |
| Exponential distribution | HPS | 66.23 ± 2.17 | 77.62 ± 1.59 | 71.70 ± 2.09 | 74.36 ± 2.07 |
| | HPS-DRS | 64.77 ± 2.14 | 79.43 ± 1.54 | 71.73 ± 2.22 | 73.65 ± 1.78 |
| Weibull distribution | HPS | 69.23 ± 3.58 | 83.11 ± 2.76 | 75.85 ± 2.95 | 80.44 ± 2.16 |
| | HPS-DRS | 72.37 ± 2.72 | 82.29 ± 3.13 | 77.17 ± 2.80 | 82.61 ± 2.75 |
| Gaussian mixture distribution | HPS | 78.47 ± 4.72 | 89.17 ± 5.29 | 83.58 ± 5.35 | 84.32 ± 2.93 |
| | HPS-DRS | **84.23 ± 3.14** | **91.50 ± 4.49** | **87.77 ± 4.62** | **91.65 ± 3.25** |

Important information of negative class may be lost, hence the performance is lowest. When increasing the value of $\alpha$, the two performances increase. When $\alpha$ is 1, over-sampling for positive class is performed without removing redundant instances for negative class. The issue of overfitting may occur due to the large amount of the positive class as well as the redundant information of negative class. Moreover, we found the G-mean and AUC to achieve the best when $\alpha$ is 65% and 60% respectively. It demonstrates that the hybrid re-sampling scheme with an appropriate re-sampling ratio can achieve optimal classification performance. Moreover, it illustrates the effectiveness of the hybrid sampling method compared with each individual sampling technique.

### 4.6. HPS-DRS vs. the state-of-the-art methods for dealing with imbalanced data

After finding out that HPS-DRS is effective on the imbalanced data classification, we also empirically test HPS-DRS against the state-of-the art methods for imbalanced data learning from algorithm level and re-sampling level, such as AdaCost [42], MetaCost [43], Tomek Link under-sampling [54], SMOTE [34] and SMOTEBoost [36]. AdaCost and MetaCost are two general cost sensitive learning integrating ensemble approaches, and the ratio cost of them are set to the reverse of the size, thus *RatioCost* = 5. All the sizes of ensemble methods are set to 50. Tomek Link is an under-sampling method; only examples belonging to the majority class are eliminated. These comparative methods are considered because they are commonly used in the research on class imbalance from the algorithm and the re-sampling perspective. SMOTEBoost are over-sampling methods based on a combination of the SMOTE algorithm with Boosting. We do not use the non-heuristic random re-sampling in our comparison since they have potential drawbacks such as information loss or causing overfitting. For all re-sampling methods, the positive class was oversampled until the classes reach a balanced distribution.

From Table 4, it is apparent that HPS-DRS achieved higher G-mean and AUC value than the other contender methods. Tomek Link performs the least since it is hard to identify the noise when the distribution is complex and imbalanced. Some border points may also be removed as noise while they are useful for training, resulting in loss of information. SMOTE is the most common technique, and many extensions of SMOTE have been proposed. However, SMOTE helps in broadening the decision region of the minority class blindly without regard to the distribution of the majority class. This leads to over-generalization so as to inevitably decrease the accuracy of the majority class. SMOTEBoost can improve the basic SMOTE since they benefit from the diversity of the ensemble framework. These re-sampling methods only consider the class skew and properties of the dataset as a whole, while datasets often exhibit characteristics and properties at a local, rather than global level. Hence, it becomes important to analyze and consider the datasets in the reduced subsets (reduced features and instances). The two general cost sensitive learning methods (AdaCost and MetaCost) are better than Tomek Link, but perform slightly worse than the other advanced re-sampling methods.It may be because the ratio misclassification cost based on the size ratio between two classes is not appropriate, resulting in obtaining an unexpected performance. The misclassification cost is vital for cost sensitive learning, and needs to be searched by some heuristic methods. The experimental results demonstrated that the Diverse Random subspace ensemble framework is an effective method achieving better performance. Moreover, HPS-DRS can improve the positive recognition ability so as to achieve excellent overall performance with G-mean and AUC.

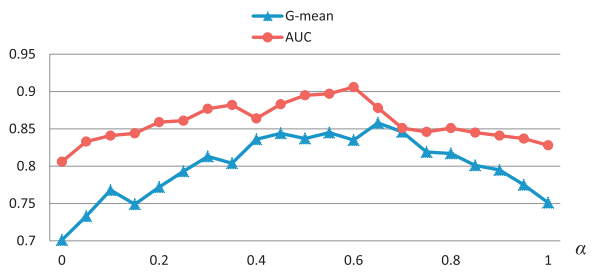



**Fig. 5.** The performance of HPS-DRS while tuning re-sampling ratio $\alpha$ in terms of G-man and AUC.

### 4.7. HPS-DRS vs. the state-of-the-art methods for false positive reduction

In this experiment, we compare our proposed method against some current methods for false positive reduction. Linear discriminant analysis (LDA) technique is most frequently used in removing false positives [11]. Random forest is an effective ensemble classification of nodule and non-nodule patterns [32]. The parameters of random forest are set as follows: the amount of trees is set to 50 and the amount of feature at each split is set to 5. Asymmetric AdaBoost (AsyAdaboost) sets different weights for two different errors [33]. The asymmetry parameter is set to 10 in the AsyAdaboost algorithm, and neural network is chosen as the base classifier.

Cost sensitive SVM (CS-SVM) is a good solution on the unbalanced data sets [31,55]. Radial basis function (RBF kernel) is a reasonable first choice for the classification of the nonlinear

**Table 4**
The comparison between our method with other approaches for imbalanced data learning.

| Method | Sensitivity (%) | Specificity (%) | G-mean (%) | AUC (%) |
|---|---|---|---|---|
| MetaCost | 77.23 ± 2.82 | 86.98 ± 2.33 | 81.90 ± 3.06 | 83.17 ± 2.15 |
| AdaCost | 81.49 ± 1.89 | 87.22 ± 2.26 | 84.25 ± 2.40 | 81.53 ± 2.04 |
| SMOTE | 78.51 ± 5.44 | 85.57 ± 4.71 | 81.93 ± 4.85 | 82.62 ± 4.21 |
| SMOTEBoost | 82.13 ± 4.50 | 87.93 ± 4.32 | 84.95 ± 4.19 | 86.30 ± 2.76 |
| Tomek link | 72.35 ± 6.11 | 85.24 ± 5.73 | 78.50 ± 5.80 | 80.78 ± 4.22 |
| HPS-DRS | **84.23 ± 3.14** | **91.50 ± 4.49** | **87.77 ± 4.62** | **91.65 ± 3.25** |

**Table 5**
The comparison between our method with other approaches for false positive reduction.

| | HPS-DRS | RF | AsyAdaboost | CS-SVM | LDA |
|---|---|---|---|---|---|
| G-mean | **87.77 ± 4.62** | 77.63 ± 3.16 | 81.37 ± 4.05 | 84.94 ± 3.87 | 72.15 ± 1.62 |
| *p*-value | – | 0.037 | 0.019 | 0.015 | 0.052 |
| AUC | **91.65 ± 3.25** | 79.22 ± 3.12 | 87.18 ± 3.44 | 88.36 ± 2.23 | 75.40 ± 1.59 |
| *p*-value | – | 0.040 | 0.022 | 0.017 | 0.033 |

datasets, as it has fewer parameters ($\gamma$). Hence we choose RBF as the kernel function of CS-SVM in this experiment. We fix $C = C$ and $C+ = C \times C_{rf}$, where $C$ and $C_{rf}$ are respectively the regularization parameter and the ratio misclassification cost factor. The ranges for C and $\gamma$ are based on a grid search for SVM parameters as recommended in [56], their ranges are $(2^{-5}, 2^{15})$ and $(2^{-15}, 2^{3})$, respectively. The ratio cost is usually suggested to be the ratio of two amounts of classes. However, our experiments show that it is not always optimal. Therefore, the ratio cost is searched iteratively to get a best one, after obtaining the optimal intrinsic parameters of SVM. The range of ratio misclassification cost factor $C_{rf}$ was empirically chosen between 1 and $100 \times N_{neg}/N_{pos}$. However, it is not feasible to use a triple circulation for optimizing the best parameters, so we optimize the best parameter pair ($C$ and $\gamma$) and the cost ratio parameter simultaneously [55]. G-mean is used to guide the search of the intrinsic parameters and ratio cost. In this work we used LibSVM as the implementation of SVM [56]. All the experiments are conducted by 10-fold cross validation.

Experiment results (Table 5) demonstrate the benefit of HPS-DRS compared to other commonly used methods for false positive reduction in Lung nodule CAD. We can see that HPS-DRS outperforms other methods in terms of AUC and G-mean. Moreover, the optimized CS-SVM obtains the second best result by searching the intrinsic parameters and ratio cost guided by the metric of G-mean. LDA performs worst compared with other sophisticated classifiers or ensemble methods. That may be because high-dimensional features make LDA overfitting and the features extracted from nodule candidates are not generally linear. Although random forest is an ensemble classifier, it can also suffer from the curse of learning from an imbalanced training data set, resulting in lower G-mean. In order to evaluate the significance of the results, paired t-tests on average G-mean were performed to compare the proposed method with the other methods. At confidence level $\alpha = 0.05$, the results in Table 5 show that the proposed method significantly outperforms the other five commonly used methods. We also showed the ROC analysis of the various approaches in Fig. 6.

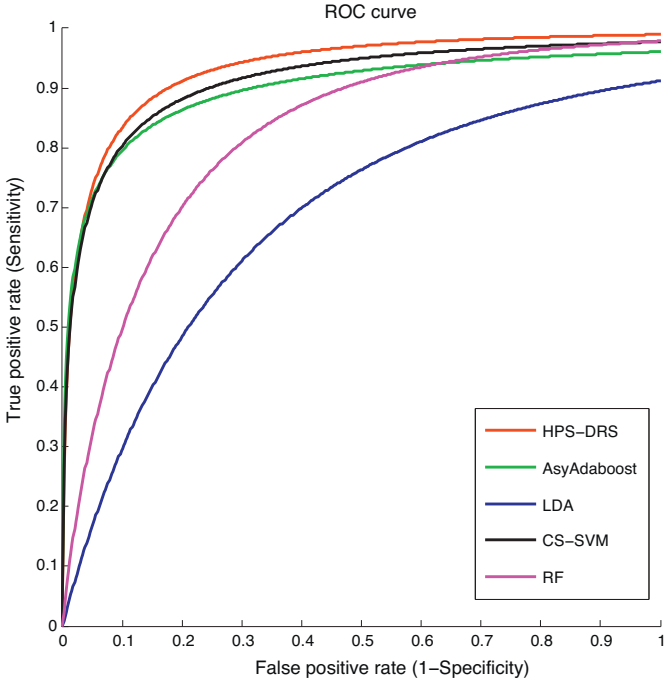



**Fig. 6.** ROC analysis of the various approaches.

### 4.8. Evaluation on LIDC database

As a further independent data set, we adopted the scans collected by the lung image database consortium (LIDC) [57], a publicly available database from the National Biomedical Imaging Archive (NBIA), and its nodules have been fully annotated by multiple radiologists. In this database, four expert chest radiologists drew outlines for nodules having effective sizes of 3 mm or greater. The ground truth was then established by a blind reading and a subsequent unblinded reading. The LIDC Database contains 399 cases, each of which includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists. Due to the limitation that our detection methods only be effective on the solid nodule, we successfully detected 209 solid nodules along with 778 non-nodules with our detection and labeling methods introduced in the first experiment using the provided ground truth information. Then the same features were computed from the VOIs to characterize the nodule candidate data (Fig. 7).

Details of the performance test for the comparative methods are shown in Table 6, which compares the performance based on the LIDC dataset with respect to sensitivity, specificity, G-mean and AUC. The proposed method reveals better performance than the other classifiers that were considered with respect to sensitivity, G-mean and AUC. From both experiments on the different lung nodule datasets, we can demonstrate this proposed approach is effective in computer aided detection of pulmonary nodules from CT scans by the re-sampling technique which can improve the quality of the data distribution of nodule candidates, combined with a random set of diverse but high performing classifiers. It could effectively reduce the number of false positives in the nodule candidates while maintaining a high sensitivity.

**Table 6**
The comparison between our method with other approaches on LIDC dataset.

| Method | Sensitivity (%) | Specificity (%) | G-mean (%) | AUC (%) |
|---|---|---|---|---|
| MetaCost | 75.47 ± 3.43 | 85.32 ± 3.17 | 80.24 ± 3.77 | 80.64 ± 2.98 |
| AdaCost | 77.76 ± 2.43 | 84.89 ± 1.87 | 81.25 ± 2.61 | 80.04 ± 3.34 |
| SMOTE | 77.58 ± 6.15 | 83.19 ± 5.83 | 80.34 ± 5.91 | 79.76 ± 5.07 |
| SMOTEBoost | 78.27 ± 5.12 | 86.44 ± 4.89 | 82.26 ± 5.46 | 82.17 ± 4.14 |
| Tomek link | 69.89 ± 6.79 | 78.60 ± 5.80 | 74.12 ± 5.94 | 76.56 ± 6.23 |
| RF | 73.61 ± 4.33 | 83.17 ± 4.10 | 78.24 ± 4.39 | 78.63 ± 4.75 |
| AsyAdaboost | 73.42 ± 5.15 | 85.54 ± 5.07 | 79.25 ± 5.21 | 81.34 ± 4.01 |
| CS-SVM | 79.90 ± 5.78 | **87.78 ± 6.14** | 83.75 ± 6.27 | 82.33 ± 5.39 |
| LDA | 72.15 ± 6.47 | 82.66 ± 4.63 | 77.23 ± 6.46 | 75.55 ± 5.17 |
| HPS-DRS | **83.17 ± 4.69** | 86.22 ± 5.02 | **84.69 ± 4.77** | **85.70 ± 4.36** |

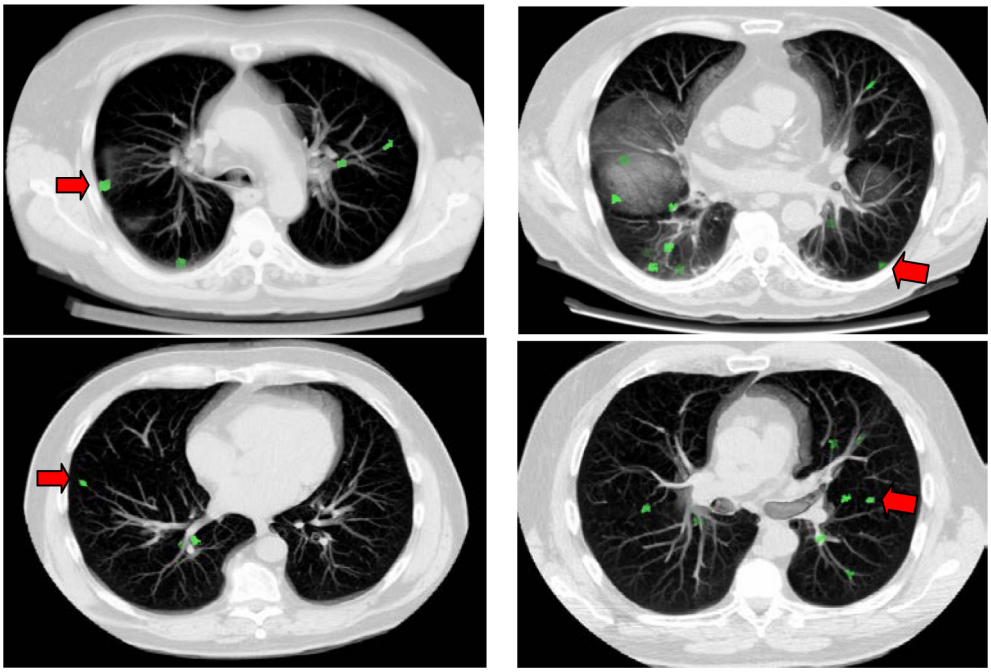


**Fig. 7.** Initial detection result of candidate nodules on LIDC. TP indicated by arrow, other spots are FP.

### 4.9. Evaluation on UCI dataset

Mining imbalanced datasets continues to be a ubiquitous problem in a large variety of applications including medicine, finance, and security. To further test the performance of our proposed method, in addition to the Lung medical database, we choose some medical datasets from public UCI datasets. The UCI Machine Learning Repository [58](http://archive.ics.uci.edu/ml/) is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. In the medical diagnosis, it is common that there is a huge disproportion in the number of cases belonging to different classes. For example, the number of cancer cases is much smaller than that of the healthy ones. Moreover, the minority class is much more important in real applications. Therefore, the data from the medical domain are mainly imbalanced data, as a dataset is said to be imbalanced if one class (negative class) outnumbers the other (positive class) and the positive class is the class of interest.

Five UCI disease datasets are chosen. Datasets which were used in these tests are in Table 7. A brief description of the datasets is given below:

**Breast-w**: Each instance represents a mass from a breast image, and features are geometric characteristics of the mass. The classification task is to determine whether the breast mass is malignant or benign.

**Pima Indian diabetes**: Each instance represents a separate member of the Pima Indian tribe, and attributes are basic medical information including age, body mass index, and blood pressure. The class attribute describes whether or not this particular individual is diabetic.

**Liver disorders**: Each instance represents a medical test on a single patient and the features are diagnostic markers that are thought to be representative of liver disease. The class indicates whether or not the patient's liver is diseased.

**SPECTF heart**: The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.

**Wisconsin Diagnostic Breast Cancer (WDBC)**: The Wisconsin Diagnostic Breast Cancer (WDBC) dataset consist of 569 instances (357 benign 212 malignant), where each one represents FNA test measurements for one diagnosis case. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

The experiments are performed using a 10-fold cross validation procedure. HPS-DRS shows appealing performances on imbalanced data through hybrid sampling with probability function and the diverse ensemble framework. HPS-DRS solves the issues of class imbalance more flexibly and efficiently. The comparative results demonstrate that our methodologies have a better predictive ability than other single and ensemble classifiers in the context of imbalanced data. From Tables 8 and 9, we can also notice that for the datasets with high dimensional features such as SPECTF and WDBC, HPS-DRS offers a great advantage over other methods. A general issue in machine learning is that using too many features in the classification algorithm can be problematic, particularly if there are irrelevant features. This can lead to overfitting, in which noise or irrelevant features may exert undue influence on the classification decisions because of the finite size of the training sample. Similar problems may arise due to redundancies in the extracted features: several features may describe the same concept. Our proposed method can mitigate the issue, and improve the generalized classification performances. However, our method cannot achieve the best result on some low dimensional datasets, such as Liver disorders which has 6 attributes. That is because the random subspace method is more effective when datasets have a large numbers

**Table 7**
UCI dataset description.

| Dataset(+) | Instances | Features | Class ratio |
|---|---|---|---|
| Breast-w (benign) | 699 | 9 | 241:458 |
| Pima Indian Diabetes (positive) | 768 | 8 | 268:500 |
| Liver disorders | 345 | 6 | 145:200 |
| SPECTF (normal) heart | 267 | 44 | 55:212 |
| WDBC (malignant) | 569 | 32 | 212:357 |

**Table 8**
The comparison between our method with other approaches on UCI dataset.

| Dataset | Metric | MetaCost | AdaCost | SMOTE | SMOTEBoost | Tomek link |
|---|---|---|---|---|---|---|
| Breast-w | G-mean | 0.942 | 0.966 | 0.970 | 0.972 | 0.964 |
| | AUC | 0.951 | 0.969 | 0.962 | 0.976 | 0.920 |
| Pima | G-mean | 0.741 | 0.738 | 0.744 | 0.741 | 0.720 |
| | AUC | 0.846 | 0.867 | 0.863 | 0.864 | 0.838 |
| Liver disorders | G-mean | 0.607 | 0.612 | 0.629 | 0.645 | 0.603 |
| | AUC | 0.803 | 0.819 | 0.826 | 0.835 | 0.794 |
| SPECTF heart | G-mean | 0.771 | 0.766 | 0.769 | 0.773 | 0.707 |
| | AUC | 0.872 | 0.887 | 0.871 | 0.874 | 0.855 |
| WDBC | G-mean | 0.910 | 0.923 | 0.951 | 0.967 | 0.917 |
| | AUC | 0.952 | 0.964 | 0.982 | **0.993** | 0.955 |

**Table 9**
The comparison between our method with other approaches on UCI dataset.

| Dataset | Metric | Random forest | AsyAdaboost | CS-SVM | LDA | HPS-DRS |
|---|---|---|---|---|---|---|
| Breast-w | G-mean | 0.972 | 0.958 | **0.988** | 0.919 | 0.976 |
| | AUC | 0.933 | 0.953 | **0.986** | 0.931 | **0.986** |
| Pima | G-mean | 0.727 | 0.752 | 0.758 | 0.708 | **0.779** |
| | AUC | 0.862 | 0.851 | 0.884 | 0.853 | 0.887 |
| Liver disorders | G-mean | 0.594 | 0.609 | **0.655** | 0.578 | 0.629 |
| | AUC | 0.782 | 0.799 | **0.837** | 0.779 | 0.831 |
| SPECTF Heart | G-mean | 0.745 | 0.763 | 0.819 | 0.705 | **0.826** |
| | AUC | 0.872 | 0.881 | **0.893** | 0.830 | 0.882 |
| WDBC | G-mean | 0.887 | 0.902 | 0.959 | 0.863 | **0.975** |
| | AUC | 0.941 | 0.944 | 0.987 | 0.922 | **0.993** |

attributes. With a very small number of attributes, each classifier receives a small set of features and thus is weak. The CS-SVM classifier performs satisfactorily for some data relatively easy to be classified. However, it cannot achieve an expected performance on some complex data, and the optimization of parameters is time consuming. Empirical evaluation on a wide variety of imbalanced datasets establishes the superiority of the new algorithm.

## 5. Conclusion

The false positive reduction is a class imbalance task in the Lung nodule detection. In this paper, we have proposed a new algorithm, hybrid probabilistic sampling method combined with weighted random subspace methods. DRS is an effective framework for imbalanced data learning as it provides varied subsets and weighted voting according to each classifier performance. Moreover, HPS can solve the between-class and within-class imbalance simultaneously. Through theoretical justifications and empirical studies, we demonstrated the effectiveness of the method on the performance of reducing false positives. The methods proposed could be applied on many other potential lesion detection, such as mass, polyp, and microcalcification. Moreover, it can also be applied to other machine learning problems such as computer-aided diagnosis. Furthermore, in this paper, our methods are only evaluated on binary class imbalanced classification. In the future, we will extend our methods to multi-class imbalanced classification.

## Acknowledgments

This work is supported by the Alberta Innovates Centre for Machine Learning as well as the National Natural Science Foundation of China (61001047) and one author was supported by the China Scholarship Council for two years at the University of Alberta.

## References

[1] Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, et al. Cancer statistics, 2005. CA: A Cancer Journal for Clinicians 2005;55(1):10–30.
[2] Cancer Facts and Figures 2012. The American Cancer Society 2012.
[3] Austin JH, Müller NL, Friedman PJ, Hansell DM, Naidich DP, Remy-Jardin M, et al. Glossary of terms for CT of the lungs: recommendations of the Nomenclature Committee of the Fleischner Society. Radiology 1996;200(2):327–31.
[4] Li Q. Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Computerized Medical Imaging and Graphics 2007;31(4–5):248–57.
[5] Temesguen M, Russell CH, Steven KR. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Medical Image Analysis 2010;14:390–406.
[6] Gomathi M, Thangaraj P. Computer aided medical diagnosis system for detection of lung cancer nodules: a survey. International Journal of Computational Intelligence Research 2009;5(4):453–62.
[7] Dhara AK, Mukhopadhyay S, Khandelwal N. Computer-aided detection and analysis of pulmonary nodule from CT images: a survey. IETE Technical Review 2012;29(4):265–75.
[8] Boroczky L, Zhao LZ, Lee KP. Feature subset selection for improving the performance of false positive reduction in lung nodule CAD. IEEE Transactions on Information Technology in Biomedicine 2006;10(3):504–11.
[9] Choi WK, Choi TS. Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. Information Sciences 2012;212(1):57–78.
[10] Suzuki K, Armato SG, Li F, Sone S, Doi K. Massive training artificial neural network for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. Medical Physics 2003;30:1602–17.
[11] Ge Z, Sahiner B, Chan HP, Hadjiiski LM, Wei J, Bogot N, et al. Computer aided detection of lung nodules: false positive reduction using a 3-D gradient field method. In: Proceedings of SPIE 5370. 2004. p. 1076–82.
[12] Yang M, Periaswamy S, Wu Y. False positive reduction in lung GGO nodule detection with 3D volume shape descriptor. In: IEEE international conference on acoustics, speech and signal processing, vol. 1. 2007. p. 437–40.
[13] Melli G, Wu X, Beinat P, Bonchi F, Cao L, Duan R, et al. Top-10 data mining case studies. International Journal of Information Technology and Decision Making 2012;11(2):389–400.
[14] Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter 2004;6(1):1–6.
[15] He H, Garcia E. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 2009;21(9):1263–84.
[16] Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Networks 2008;21(2–3):427–36.
[17] Rao RB, Fung G, Krishnapuram B, Bi J, Dundar M, Raykar V, et al. Mining medical images. In: Proceedings of the third workshop on data mining case studies and practice prize, fifteenth annual SIGKDD international conference on knowledge discovery and data mining. 2009.
[18] Ren J. ANN vs. SVM: which one performs better in classification of MCCS in mammogram imaging. Knowledge-Based Systems 2012;26:144–53.
[19] Yang X, Zheng Y, Siddique M, Beddoe G. Learning from imbalanced data: a comparative study for colon CAD. In: Proceedings of the SPIE. 2008. p. 6915.
[20] Malof JM, Mazurowski MA, Tourassi GD. The effect of class imbalance on case selection for case-based classifiers: an empirical study in the context of medical decision support. Neural Networks 2012;25:141–5.

[21] Yang Q, Wu X. 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making 2006;5(4):597–604.
[22] Weiss G. The impact of small disjuncts on classifier learning. Annals of Information Systems 2010;5(8):193–226.
[23] Jo T, Japkowicz N. Class imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter 2004;6(1):40–9.
[24] Holte RC, Acker LE, Porter BW. Concept learning and the problem of small disjuncts. In: Proceedings of the 7th international joint conference on artificial intelligence. 1989. p. 813–8.
[25] Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics 2010;11(1):523–41.
[26] Ho T. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998;20(8):832–44.
[27] Li Q, Doi K. Analysis and minimization of overtraining effect in rule-based classifiers for computer-aided diagnosis. Medical Physics 2006;33:320–8.
[28] Tao Y, Lu L, Dewan H, Chen A, Corso J, Xuan J, et al. Multi-level ground glass nodule detection and segmentation in CT lung images. In: Medical image computing and computer-assisted intervention-MICCAI. 2009. p. 715–23.
[29] Jiang J, Trundle P, Ren J. Medical image analysis with artificial neural networks. Computerized Medical Imaging and Graphics 2010;34(8):617–31.
[30] Wang Q, Kang W, Wu C, Wang B. Computer-aided detection of lung nodules by SVM based on 3D matrix patterns. Clinical Imaging 2013;37(1):62–9.
[31] Campadelli P, Casiraghi E, Valentini G. Support vector machines for candidate nodules classification. Neurocomputing 2005;68:281–8.
[32] Lee SLA, Kouzani AZ, Hu EJ. Random forest based lung nodule classification aided by clustering. Computerized Medical Imaging and Graphics 2010;34(7):535–42.
[33] Dolejsi M, Kybic J, Tuma S, Polovincák M. Reducing false positive responses in lung nodule detector system by asymmetric Adaboost newblock. In: Proceedings of fifth IEEE international symposium on biomedical imaging ISBI. 2008. p. 656–9.
[34] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002;16(1):321–57.
[35] Ramentol E, Caballero Y, Bello R, Herrera F. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. Knowledge and Information Systems 2012;33(2):245–65.
[36] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: improving prediction of the minority class in boosting. In: Knowledge discovery in databases: PKDD. 2003. p. 107–19.
[37] Wang BX, Japkowicz N. Boosting support vector machines for imbalanced data sets. Knowledge and Information Systems 2010;25(1):1–20.
[38] Sun Y, Kamel MS, Wang Y. Boosting for learning multiple classes with imbalanced class distribution. In: Proceedings of the sixth international conference on data mining. 2006. p. 592–602.
[39] Zhou ZH, Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 2006;18(1):63–77.
[40] Galar M, Fernandez A, Barrenechea E, Bustince H, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C 2012;42:463–84.
[41] Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 2009;39(2):539–50.
[42] Fan W, Stolfo SJ, Zhang J, Chan PK. AdaCost: misclassification cost-sensitive boosting. In: Proceedings of international conference machine learning. 1999. p. 97–105.
[43] Domingos P. MetaCost: a general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. 1999. p. 155–64.
[44] Figueiredo MAT, Jain AK, Doi K. Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 2002;24(3):381–96.
[45] Denil M, Trappenberg T. Overlap versus imbalance. In: Proceedings of the 23rd Canadian conference on advances in artificial intelligence. 2010. p. 220–31.
[46] Li Q, Sone S, Doi K. Selective enhancement filters for nodules, vessels, and airway walls in two-and three-dimensional CT scans. Medical Physics 2003;30:2040–51.
[47] Li Q, Li F, Doi K. Computerized detection of lung nodules in thin-section CT images by use of selective enhancement filters and an automated rule-based classifier. Academic Radiology 2008;15(2):165–75.
[48] Lu X, Wei GQ, Qian J, Jain AK. Learning-based pulmonary nodule detection from multislice CT data. In: Proceedings of CARS 2004 computer assisted radiology and surgery. 2004. p. 1356–63.
[49] Yoshida H, Nappi J. Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps. IEEE Transactions on Medical Imaging 2001;20(12):1261–74.
[50] Koenderink JJ. Solid shape. Cambridge: Cambridge University Press; 1990.
[51] Kawata Y, Niki H, Ohmatsu H, Kakinuma R, Eguchi K, Kaenko M, et al. Quantitative surface characterization of pulmonary nodules based on thin-section CT images. IEEE Transactions on Nuclear Science 1998;45(4):2132–8.
[52] Kobatake H, Murakami M. Adaptive filter to detect rounded convex regions: Iris filter. In: Proceedings of the 13th international conference on pattern recognition. 1996. p. 340–4.
[53] Wu WJ, Lin SWH, Moon WK. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. Computerized Medical Imaging and Graphics 2012;36:627–33.
[54] Tomek I. Two modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics 1976;6:769–72.
[55] Cao P, Zhao DZ, Zaiane O. An optimized cost-sensitive SVM for imbalanced data learning. In: Proceedings of the 17th Pacific-Asia conference on knowledge discovery and data mining. 2013. p. 280–92.
[56] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification; 2003.
[57] McNitt-Gray MF, Armato SG, Meyer CR, Reeves AP, et al. The lung image database consortium (LIDC) data collection process for nodule detection and annotation. Academic Radiology 2007;14:1464–74.
[58] Asuncion G, Newman D. UCI machine learning repository; 2007.

**Peng Cao**is a PhD student enrolled in the Northeastern University, China. Peng has been working with the Alberta Innovates Center for Machine Learning (AICML) – computing science at the University of Alberta, since 2011. Peng received an MSc in computer science from Northeastern University, China, in 2005. His research interests are imbalanced data learning, medical image mining.

**Jinzhu Yang** received his PhD in computer science from Northeastern University, in 2007. Jinzhu is an associate professor in the College of Information Science and Engineering, Northeastern University.

**Wei Li** received his Master in computer science from Northeastern University, in 2006. Wei is an associate director in Key Laboratory of Medical Image Computing of the Chinese Ministry of Education.

**Dazhe Zhao** received her PhD in computer science from Berlin Institute of Technology, Germany, in 1996. Dazhe is a professor in the College of Information Science and Engineering, Northeastern University, and the director in Key Laboratory of Medical Image Computing of the Chinese Ministry of Education. She is also the president of the Research Institute of Neusoft Corporation.

**Osmar Zaiane** received his PhD in computing science from Simon Fraser University, Canada, in 1999 specializing in data mining. He is a professor at the University of Alberta, Canada, with research interest in novel data mining techniques and currently focuses on social network analysis as well as health informatics applications. He is the scientific director of AICML (Alberta Innovates Centre for Machine Learning). He regularly serves on the program committees of international conferences in the field of knowledge discovery and data mining and was the program co-chair for the IEEE international conference on data mining ICDM'2007, and the general chair for ICDM'2011. He is the associate editor of knowledge and information systems and data mining and knowledge discovery.