IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 31, 2018, accepted January 25, 2019, date of publication January 30, 2019, date of current version February 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896409

WGAN-Based Synthetic Minority Over-Sampling Technique: Improving Semantic Fine-Grained Classification for Lung Nodules in CT Images

QINGFENG WANG[1,2], XUEHAI ZHOU[1], CHAO WANG[1], ZHIQIN LIU[2], JUN HUANG[2], YING ZHOU[3], CHANGLONG LI[1], HANG ZHUANG[1], AND JIE-ZHI CHENG[4]

[1]School of Software Engineering, University of Science and Technology of China, Hefei 230026, China
[2]School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China
[3]Radiology Department, Mianyang Central Hospital, Mianyang 621000, China
[4]Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200232, China

Corresponding author: Jie-Zhi Cheng (jzcheng@ntu.edu.tw)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700900, in part by the National Natural Science Foundation of China under Grant 61772482, Grant 61501305, and Grant 61672438, in part by the Youth Innovation Promotion Association CAS under Grant 2017497, in part by the Anhui Provincial Natural Science Foundation under Grant BJ2150110002, in part by the Sichuan Provincial Open Foundation of Civil-Military Integration Research Institute under Grant 2017SCII0219 and Grant 2017SCII0220, and in part by the Key Project of Sichuan Provincial Science and Technology Innovation under Grant 19MZGC0123.

ABSTRACT Data imbalance issue generally exists in most medical image analysis problems and maybe getting important with the popularization of data-hungry deep learning paradigms. We explore the cutting-edge Wasserstein generative adversarial networks (WGANs) to address the data imbalance problem with oversampling on the minority classes. The WGAN can estimate the underlying distribution of a minority class to synthesize more plausible and helpful samples for the classification model. In this paper, the WGAN-based over-sampling technique is applied to augment the data to balance for the fine-grained classification of seven semantic attributes of lung nodules in computed tomography images. The fine-grained classification is carried out with a normal convolutional neural network (CNN). To further illustrate the efficacy of the WGAN-based over-sampling technique, the conventional data augmentation method commonly used in many deep learning works, the generative adversarial networks (GANs), and the deep convolutional generative adversarial networks (DCGANs) are implemented for comparison. The whole schemes of the minority oversampling and fine-grained classification are tested with the public lung imaging database consortium dataset. The experimental results suggest that the WGAN-based oversampling technique can synthesize helpful samples for the minority classes to assist the training of the CNN model and to boost the fine-grained classification performance better than the conventional data augmentation method and the two schemes of the GAN and DCGAN techniques do. It may thus suggest that the WGAN technique offers an alternative methodological option for the further deep learning on imbalanced classification studies.

INDEX TERMS Computer-aided diagnosis (CAD), lung nodule, computed tomography (CT), synthetic minority over-sampling, deep learning, data imbalance, adversarial neural networks.

I. INTRODUCTION

Recent advance of deep learning techniques has been shown to effectively address many medical image analysis problems like segmentation [1]–[5], lesion detection [6]–[8], differential diagnosis [9]–[13], quality assessment [14], reference plane retrieval [15], etc., with perceivable performance improvement. The deep learning techniques are equipped with the advantages of automatic feature learning, end-to-end training, etc., and thus the steps of explicit feature engineering as well as other inter-mediate processing in the conventional pattern recognition framework can be circumvented. Therefore, the performance tuning of the deep learning techniques can be relatively simple and easy. In particular, the powerful discriminative capability of the deep learning techniques may also shed a light on fine-grained medical

The associate editor coordinating the review of this manuscript and approving it for publication was Chang-Tsun Li.

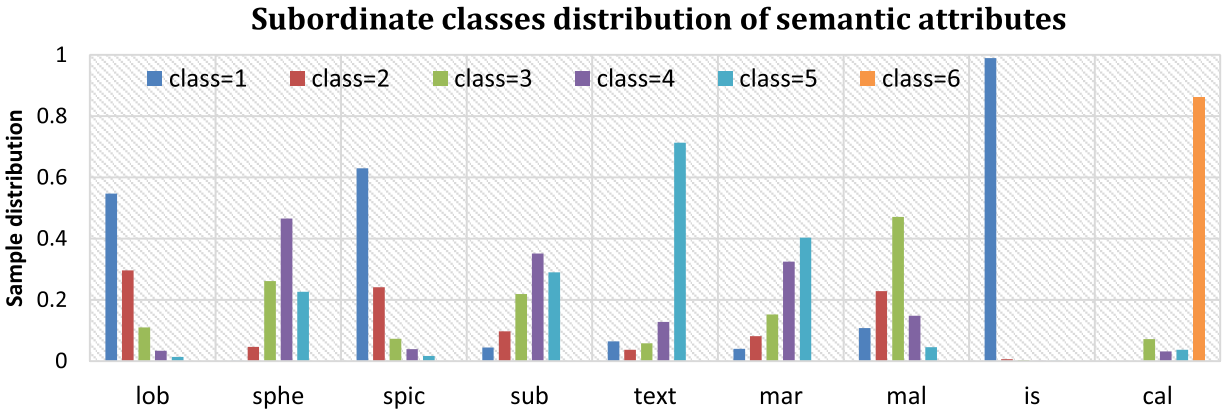


FIGURE 1. Subordinate classes distributions over all semantic attributes of nodules in LIDC dataset. "lob", "sphe", "spic", "sub", "text", "mar", "mal", "is", and "cal" are the abbreviations of "lobulation", "sphericity", "spiculation", "subtlety", "texture", "margin", "malignancy", "internal structure" and "calcification", respectively.

image analysis to attain more precise diagnosis, prognosis and prediction [16], [17]. For example, the fine-grained medical image analysis may help to achieve more accurate computerized retrieval of relevant cases [18], lesion subtype categorization [19], etc.

The main purpose of the fine-grained classification is to differentiate subordinate categories of the same base classes. The subordinate categories share very similar properties of the same base class, whereas the differences in-between the subordinate categories can be subtle. Therefore, the task of the fine-grained classification can be very challenging. The fine-grained classification is particular of help for the applications like online shopping recommendation system, etc. In the medical context, the fine-grained image analysis problem has been less explored. The exploration of the fine-grained medical image analysis is limited by the available data and annotation. In particular, the data annotation requires professional knowledge and the annotation cost can be very expensive. Meanwhile, the data of the different subordinate categories can also be very imbalanced and thus impose more difficulty on the fine-grained medical image analysis.

In recent years, there are few studies elaborating on the computerized fine-grained analysis for medical images. Zhang *et al.* [18] developed a template matching framework to perform the fine-grained differentiation of the two types of lung cancers, i.e., adenocarcinoma and squamous carcinoma, in histological images. The work [18] requires large scale segmented cancer cells as templates to achieve promising performance. For the lung nodule analysis in CT images, Chen *et al.* [20], [21] recently leverage the deep learning techniques of the convolutional neural network (CNN) and stacked denoising autoencoder (SDAE) and multi-task learning technique to attain fine-grained semantic attributing of pulmonary nodules. Specifically, a pulmonary nodule can be profiled with 9 semantic medical descriptive terms like spiculation, lobulation, subtlety, etc. For each semantic term, there are around 5 to 6 subordinate scoring classes to suggest the degree or instantiation of the corresponding term. The main idea of the works [20], [21] lies to use deep learning techniques for the learning of useful features and employs multi-task learning to explore sharable and term-specific features to attain satisfactory fine-grained semantic attributing performance on the public Lung Image Database Consortium (LIDC) dataset [22], which contains at least 1010 CT scans from 1010 patients. However, since the distributions of the subordinate classes of each semantic term can be very skewed as shown in Fig. 1, the performance of the deep features learning and multi-task framework was limited by the data imbalance issue.

By and large, the data imbalance issue generally exists in most medical image analysis problems. It is because that the number of cases with diseases is relatively smaller than the number of normal cases. Meanwhile, the case distribution subordinate types of one specific disease can be very skewed due to the factors of race, gender, disease rarity, and so on. For example, Fig. 1 illustrates the distributions of the subordinate classes of the 9 semantic terms for nodules in the LIDC dataset. In Fig. 1, "lob", "sphe", "spic", "sub", "text", "mar", "mal", "is", and "cal" are the abbreviations of the terms "lobulation", "sphericity", "spiculation", "subtlety", "texture", "margin", "malignancy", "internal structure" and "calcification", respectively. These semantic terms can be commonly found in the radiology reports to describe the semantic characteristics of the nodules for the diagnostic reference. The terms "subtlety" and "sphericity" suggest if the nodule is easy to identify and the roundness of nodule shape, respectively. The term "margin" describes how well-defined of nodule margin is, whereas the "lobulation" and "spiculation" terms suggest if the nodule has lobulation or spiculation in shape, respectively. The term "texture" indicates if the nodule appears solid in the image, while the term "malignancy" is the subjective assessment of the malignancy likeliness by the radiology. The term "internal structure" specifies the nodule internal can be soft tissue, fluid, fat or air. The term "calcification" stands for the calcification pattern of the nodules. More details can be found in [23].

Referring to Fig. 1, it can be observed that the distributions of all subordinate classes for all 9 semantic terms are extremely imbalanced. For some subordinate classes, the sample numbers are significantly less than the sample numbers of some other subordinate classes. In some cases, the sample number of the majority subordinate class is nearly 20 times greater than the sample numbers of the minority subordinate classes. Consequently, the data imbalance issue in the fine-grained subordinate classes can easily bias the learning frameworks, but sadly was not elaborated in previous works.

In this study, the data imbalance issue of medical image data is explicitly addressed and tested on the LIDC dataset. The specific approach in this study is to explore the data synthesis to augment the sample numbers in the minority classes. The commonly-used conventional data augmentation techniques may involve random image translation, rotation, flipping w.r.t. horizontal or vertical direction, etc. Since the conventional data augmentation techniques don't consider the data distributions of classes, the efficacy of over-sampling for the minority classes may be limited for the data with extreme imbalanced distribution. In this paper, we investigate the deep learning approach called generative adversarial networks (GAN) [24] to synthesize samples within general distribution-aware decision region for the minority classes to combat multi-class fine-grained data imbalance problem.

The GAN technique was firstly introduced by Goodfellow *et al.* [24] and basically is constituted of two networks of a generator and a discriminator. The two networks are trained at the same time and compete against each other in a minimax game. The generator is trained to fool the discriminator by synthesizing realistic samples, whereas the discriminator is trained to be equipped high discriminative capability for the synthetic samples. However, the training of the two networks can be quite unstable and may suffer mode collapse problem [24]. Therefore, the synthesized samples by the generator can be easily noisy and incomprehensible. To further improve the capability of the generator, the technique of the deep convolutional generative adversarial networks (DCGAN) [25] was proposed by imposing a set of constraints on the architectural topology of GAN to stabilize the training process. Although better synthetic quality can be achieved with the DCGAN technique, the training of both GAN and DCGAN share the same problem of not easy to reach convergence. Therefore, to train a promising GAN and DCGAN can be very difficult.

To alleviate the issue of training difficulty, the Wasserstein GAN [26], denoted as WGAN for short, was developed by employing the Earth Mover (EM) distance for better measurement of the distances between distributions. With the EM distance, the WGAN is equipped with better converge capability and the training of the WGAN can be more stable and better withstand the problems like mode collapse, etc.

The family of the GAN techniques has been shown successfully in the applications of text-to-image [27], [28] as well as image-to-image translation [29], image super-resolution [30], etc. In medical image analysis, the GAN technique was also introduced to synthesis images of different modality [31], [32], denoising for low dose CT [33], segmentation [34], image reconstruction [35], etc. The major purpose of using GAN is to approximate the object distribution for better performance in each specific application. To our best knowledge, the GAN technique has been less exploited to approach the data imbalance issue in the domain of medical image analysis. In this paper, we adopt the WGAN for the over-sampling of the minority classes. Through the adversarial iterative training, the distribution of the synthesized images will approximate the distribution of the authentic CT images. Fig. 2 compares the authentic samples of the subordinate classes of the ''texture'' attribute with the synthesized samples by the WGAN. The majority class of the ''texture'' attribute is the class 5. Therefore, no over-sampling is performed for this class. As can be found, the synthesized samples are quite similar but different to the authentic samples. The process of synthesizing samples with WGAN considers the distributions of classes, and, thus the synthesized CT image samples may reserve more general distribution-aware decision regions for classifier than conventional approaches. Meanwhile, the difficult of the fine-grained classification can also be observed in Fig. 2. The classification between the two consecutive subordinate classes is quite challenging.

The contribution of this work can be summarized in twofold. First, the WGAN-based synthetic over-sampling technique is presented for the data augmentation of the minority classes to tackle the imbalance issue. The WGAN technique attempts to synthesize samples by the approximation of the original data distributions of the minority classes. Second, our method is also applied to the challenging problem of the fine-grained classification on the 7 semantic attributes of the LIDC lung nodules. It will be shown that the WGAN-based synthetic over-sampling technique can improve the fine-grained classifications on highly imbalanced medical data and provide better and useful synthesized minority class samples than those transformed samples with the conventional data augmentation method, which is commonly used in many deep learning paradigms. It is worth noting that the goal of this study is to illustrate the efficacy of the WGAN-based synthetic over-sampling technique on the fine-grained classification for the application with extreme data imbalance. Therefore, we don't formulate the semantic attributing of the lung nodules into a regression framework as shown in [20] and [21]. Since the studies don't consider the data imbalance issue, the optimization of regression may easily favor the majority class with smaller regression error and scarified the accuracy of the minority classes.

II. MATERIALS AND METHODS

A. DATASET

The LIDC dataset includes more than 1010 thoracic CT scans from 1010 patients, where each scan was reviewed and annotated by four experienced thoracic radiologists with rigorous reading protocol. In total, 2632 nodules in the LIDC

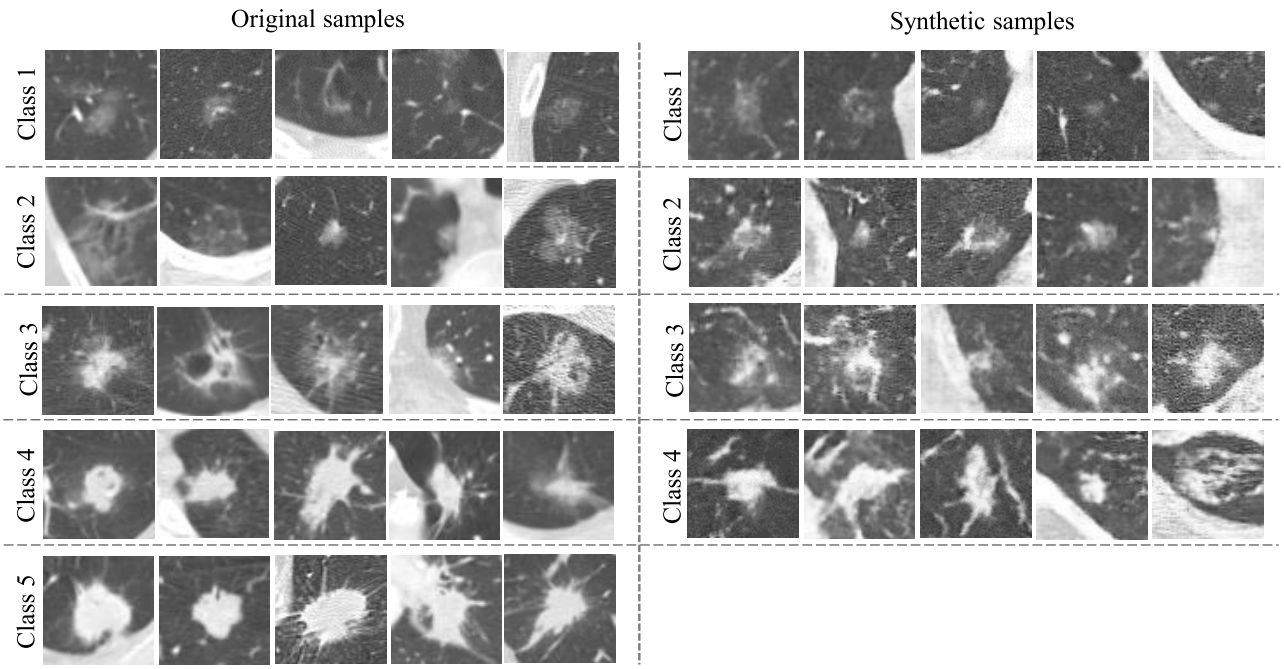


FIGURE 2. Illustration of all subordinate classes in the attribute "texture". The images shown in the left part are the authentic samples, whereas the synthetic samples by WGAN are shown in the right part.

dataset are involved in [36]. The region of interests (ROIs) in the slices that depict each nodule is cropped into 64 × 64 pixels and normalized with the lung HU window range level of [−1400, 200]. Referring to [36] and [37], the size of the largest nodule in the LIDC dataset in the transversal CT slice is no more than 64 × 64 pixels. Therefore, the setting of the ROI size can sufficiently enclose all nodules in the LIDC dataset. Each nodule was annotated with 9 semantic attribute scores by at least one radiologist. If one nodule was annotated by more than one radiologist, the semantic attribute scores from all radiologists are averaged as representative scores for training and testing [20], [21]. For the robustness of the performance evaluation, 5-fold cross validation scheme based on nodule unit is implemented for both the data augmentation step and the fine-grained classification of each semantic attribute.

In this study, the seven semantic attributes shown in Table 1 are adopted to illustrate the efficacy of the over-sampling technique with WGAN. The ''is'' and ''cal'' attributes are excluded as the sample distributions of the two attributes are too skewed to be processed. The class distributions of ''is'' and ''cal'' are (2606/15/8/2/1) and (0/0/189/82/97/2264) respectively. On the other hand, the original number of the subordinate classes of the rest seven attributes is 5. However, we found that the sample number of the subordinate class 1 in ''sphe'', subordinate class 5 in ''lob'' and subordinate class 5 in ''spic'' are 2, 36 and 44 respectively. By comparing to the number of samples in the majority class w.r.t. the above classes, the sample numbers are too small to fit the scheme of the 5-fold cross validation. Therefore, these minority subordinate classes are merged into their neighboring classes. Specifically, the class with score 1 in the attribute ''sphe'' is merged into the class with score 2, and the classes with score 5 in both attributes ''lob'' and ''spic'' are merged into the class with score 4, respectively. After the merging process, the numbers of the subordinate classes of the attributes ''sphe'', ''lob'' and ''spic'' are 4, see Table 1. Here, the majority class is defined as the class with the largest samples. The imbalanced ratio throughout this paper is defined as the ratio of the sample numbers of the majority class to the minority class [38], [39]. As can be found in Table 1, even with the merge preprocessing, the largest imbalanced ratios of all seven semantic attributes range from 8.03 to 19.15, suggesting that **the class imbalance issue remains very serious**. In this study, the schemes of GAN, DCGAN and WGAN synthetic techniques and a conventional data augementation method are employed to augment samples of the minority subordinate classes to improve the sementic fine-grained classification of lung nodules in CT images.

TABLE 1. Number of subordinate classes (#), sample distribution and imbalanced ratios for each semantic attribute.

Attribute	#	Distribution	Imbalanced Ratio
lob	4	**1439**/780/289/124	1.0/1.84/4.98/**11.60**
sphe	4	124/687/**1226**/595	**9.89**/1.78/1.0/2.06
spic	4	**1658**/636/192/146	1.0/2.61/8.64/**11.36**
sub	5	115/255/576/**924**/762	**8.03**/3.62/1.60/1.0/1.21
text	5	169/98/152/336/**1877**	11.11/**19.15**/12.35/5.59/1.0
mar	5	104/214/400/854/**1060**	**10.19**/4.95/2.65/1.24/1.0
mal	5	283/600/**1240**/390/119	4.38/2.07/1.0/3.08/**10.42**

B. OVER-SAMPLING FOR MINORITY SUBORDINATE CLASSES IN SEMANTIC ATTRIBUTES

In this study, the technique of the generative adversarial networks (GAN) is employed for the over-sampling of the

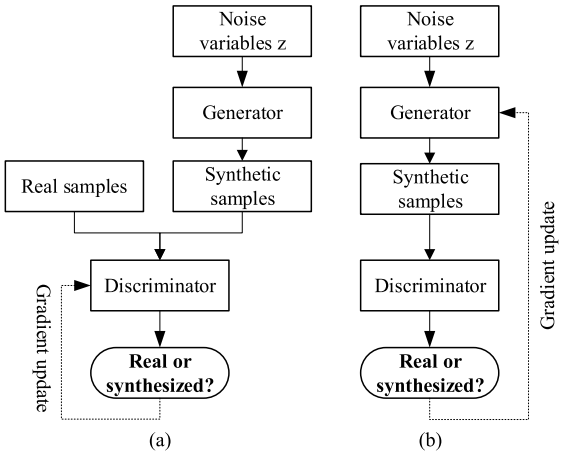


FIGURE 3. The training flowchart of a typical GAN. (a) Training path of the discriminator network; (b) training path of the generator network.

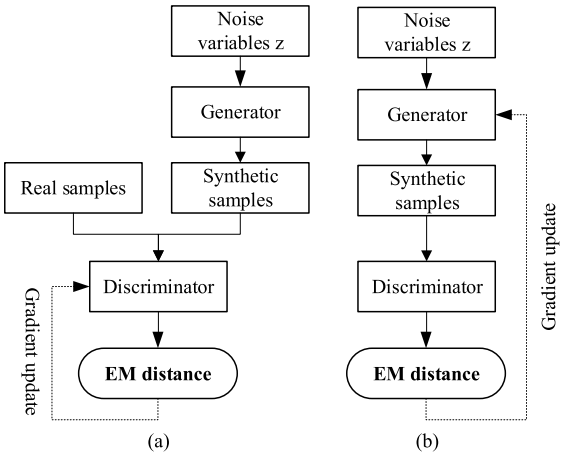


FIGURE 4. The training flowchart of WGAN. (a) Training path of the discriminator network; (b) training path of the generator network.

minority subordinate classes. Typical GAN can be constituted of a generator and a discriminator network. The functionality of the generator network is to synthesize samples by estimating the underlying distribution of the target domain, whereas the discriminator aims to differentiate the true samples and the synthetic samples derived from the generator. The optimization process of the GAN pushes the generator to synthesize plausible samples that can fool the discriminator, while also sharpens the differentiation capability of the discriminator. Therefore, the effectively training of the GAN needs to optimize two networks. The concept of the typical GAN is illustrated in Fig. 3.

Since it needs to train the networks of the generator and discriminator for the GAN, the optimization process can be difficult and may suffer several drawbacks. First, the training of the GAN is relatively unstable and may highly depend on the competition between the generator and the discriminator within the minimax game framework [24]. In other words, a good equilibrium between the generator and the discriminator is important to yield good quality of sample synthesis. However, the gradient descent during the training of the networks can't always promise a good equilibrium. For example, if a discriminator is equipped with high differentiation capability in the training process, the generator's gradient may vanish quickly. Therefore, the optimization of the generator may not be able to proceed to approximate the true distribution of the target domain. As suggested in [40], a better generator can be trained if the discriminator is deliberately weaken. In such case, it may then require several passes of trial and error and turn the whole training process difficult and unstable. Second, there exists the so-called mode collapse in GAN, where the generator tends to produce samples with low variety. The mode collapse is caused by the cases that the generator is trapped to the same local minimum of the cost function to synthesize similar samples. In such case, the generated samples are not sufficiently diverse to represent the whole distribution of the target domain, and hence is not helpful to address the problem of the data imbalance.

The Wasserstein GAN (WGAN) is a new GAN to ease the training difficulty of the typical GAN and avoid the potential problem of the mode collapse. The overview of the workflow structure of the WGAN for samples synthesis is shown in Fig. 4. Comparing to Fig. 3, the objective functions of the discriminator between the WGAN and the typical GAN are different. For the typical GAN, the objective function of the discriminator is determined with the binary classification of true and synthesized samples, whereas the objective function of the WGAN's discriminator is represented by the Earth-Mover (EM) distance between real and synthesized distributions. To this end, the learning of the WGAN's discriminator is formulated as a regression task but not classification.

The incorporation of the Earth Mover (EM) distance for the measurement of the two comparing distributions in the WGAN can avoid the asymmetry problem of the Kullback-Leibler divergence [40] that could lead to mode collapse, as well as the discontinuous issue of the loss function with Jensen-Shannon divergence that may result in unsatisfactory synthetic results. The EM distance can provide reliable and usable gradient for the loss function to more easily achieve synthesis results with better quality. The EM distance between the real samples' distribution $\mathbb{P}_r$ and the synthetic samples' distribution $\mathbb{P}_s$ can be defined as

$$W(\mathbb{P}_r, \mathbb{P}_s) = \inf_{\delta \in \prod(\mathbb{P}_r, \mathbb{P}_s)} \mathbb{E}_{(x,y)\sim\delta} \parallel x - y \parallel, \tag{1}$$

where x and y stand for the real and the synthetic samples, respectively, and $\prod(\mathbb{P}_r, \mathbb{P}_s)$ suggests the set of all joint distribution $\delta(x, y)$, where the marginal distributions of x and y are $\mathbb{P}_r$ and $\mathbb{P}_s$, respectively. Intuitively, the EM distance can be interpreted as the minimal transported "mass" from y to x for the purpose of transforming the distribution $\mathbb{P}_s$ to the distribution $\mathbb{P}_r$.

However, the infimum in e.q. (1) is highly intractable. Instead, referring to [26], the solving of e.q. (1) can be sought by

$$\max_D \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{y\sim\mathbb{P}_s}[D(y)], \tag{2}$$

where D stands for the neural network of the discriminator. The EM distance can then be approximately sought with the optimization of the discriminator D, which is driven by maximization the term $\mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{y\sim\mathbb{P}_s}[D(y)]$ [26]. The generator network, G, aims to synthesize samples with the distribution $\mathbb{P}_s$ that approximates the real data distribution $\mathbb{P}_r$. During the training process, the generator network will map a random vector z, which is commonly drawn from a normal distribution $p(z)$. By considering the sample synthesis by the generator G, the EM distance can be rewritten as

$$\max_D \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{z\sim p(z)}[D(G(z))]. \tag{3}$$

Referring to equations (1-3), the solving of the infimum in e.q. (1) was equivalently transformed by approximately seeking the maximum in e.q. (3). The maximum in e.q. (3) can suggest the EM distance between $\mathbb{P}_r$ and $\mathbb{P}_s$. Furthermore, we hope to change $\mathbb{P}_s$ to close to $\mathbb{P}_r$ as much as possible. This can be determined by adjusting the synaptic weights of the generator G, which is equivalent to $\min_G W(\mathbb{P}_r, \mathbb{P}_s)$. Therefore, the whole training process of the WGAN can then be expressed as

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{z\sim p(z)}[D(G(z))]. \tag{4}$$

Accordingly, the training of the WGAN can then be interpreted as a two-player minimax game between the discriminator D and the generator G, and then can be achieved by iteratively optimizing the discriminator D by

$$\max \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{z\sim p(z)}[D(G(z))], \tag{5}$$

as well as seeking the better generator that satisfies

$$\min -\mathbb{E}_{z\sim p(z)}[D(G(z))]. \tag{6}$$

The networks of the discriminator and the generator can be both CNNs. The architectures of the discriminator and the generator of the WGAN in this study are shown in Fig. 5. Specifically, the input layer of the discriminator is an image sample (either real or synthetic) with dimensions of 64 × 64 pixels. The following layers of the discriminator are a series of convolutional layers paired with batch normalization and leaky rectified linear unit (LReLU). Batch normalization can stabilize the leaning process and enable the gradient flow toward deeper layers. Referring to [25], the batch normalization is not recommended to be implemented for the input layer of the discriminator and the output layer of the generator to avoid sample oscillation and model instability.

The generator is structured to output the samples with the same dimensionality of inputs for the discriminator. The input of the generator is a random vector with 100 dimensions initialized from a normal distribution. The random noise vector is reshaped to 100 × 1 × 1 and then filtered with transposed convolution, denoted as CONVT, to 512 channels 4 × 4 feature maps. The CONVT is also named as deconvolution [25]. The following four layers of the generator are also CONVT layers. For the generator, rectified linear unit (ReLU) is used as the activation function of the neurons whereas the LReLU

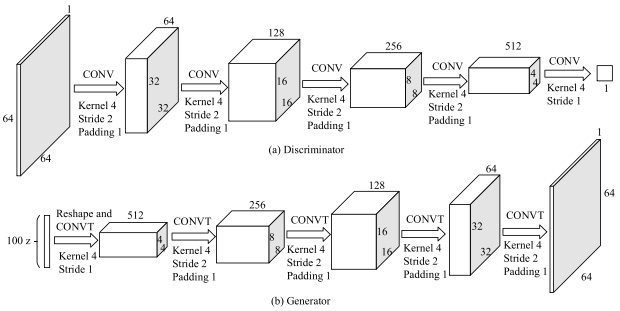


FIGURE 5. The architecture of our WGAN. (a) Discriminator architecture; (b) generator architecture.

TABLE 2. The Architecture of the fine-grained classification CNN model. The abbreviations of "C", "M", and "FC" stand for convolution, max pooling and full connected layer, respectively. The "K" is the number of the subordinate classes of each semantic attribute.

Layer	Feature maps	kernel size	Stride
input	64x64x1	-	-
C1	60x60x20	5	1
M1	30x30x20	2	2
C2	26x26x50	5	1
M2	13x13x50	2	2
C3	9x9x100	5	1
M3	5x5x100	2	2
FC4	500	-	-
FC5	K	-	-

is adopted for the neurons in the discriminator network. Batch normalization is also implemented for the stabilization of the learning process. The four CONVT layers subsequently double the dimensions of the feature maps, whereas the number of the channels is sequentially halved. The output layer of the generator is constituted with 64 × 64 pixels. The optimization of the WGAN is carried out with the RMSProp algorithm. The generator and the discriminator networks of the WGAN are initialized from scratch (random initialization from zero-centered normal distribution with standard deviation 0.02). The slope of the leak for the LReLU in the discriminator network is set to 0.2, whereas the learning rates for the both discriminator and generator are 5e-5.

C. FINE-GRAINED CLASSIFICATION FOR THE SEMANTIC ATTRIBUTES

The major purpose of this study is to illustrate the over-sampling effectiveness with the WGAN. Therefore, a normal CNN model is employed for the task of the fine-grained classification for each semantic attributes of lung nodules. The detailed architecture configuration of the category model is described in Table 2. The optimization of the CNN is sought by the stochastic gradient descent (SGD) algorithm and the momentum parameter is set as 0.9. The learning rate is set to be 0.0001 and the number of the training epoches is 100. The batch size of the training is 64 and the weight decay is set as 0.0005. The architecture of the fine-grained classification CNN model is extended from the standard shallow convolutional neural network LeNet. The hyper-parameters such as momentum, batch size and weight decay are also set as the default values used in the LeNet. The learning rate and the

TABLE 3. The Data partitions in the 5-fold CV w.r.t the 7 semantic attributes. "class1", "class2", "class3", "class4" and "class5" are abbreviated as "c1", "c2", "c3", "c4" and "c5", respectively.

Attribute	Fold	Subordinate classes distribution		Batch size	
		Training samples	Validation samples	WGAN	GAN and DCGAN
lob		c1/c2/c3/c4	c1/c2/c3/c4	c1/c2/c3/c4	c1/c2/c3/c4
	1,2,3,4	1151/624/231/99	288/156/58/25	384/312/231/99	144/156/116/99
	5	1152/624/232/100	287/156/57/24	384/312/232/100	144/156/116/100
sphe		c1/c2/c3/c4	c1/c2/c3/c4	c1/c2/c3/c4	c1/c2/c3/c4
	1,2,3	99/550/981/476	25/137/245/119	99/275/327/238	99/110/123/119
	4	99/549/981/476	25/138/245/119	99/275/327/238	99/110/123/119
	5	100/549/980/476	24/138/246/119	100/275/327/238	100/110/123/119
spic		c1/c2/c3/c4	c1/c2/c3/c4	c1/c2/c3/c4	c1/c2/c3/c4
	1,2,3,4	1326/509/154/117	332/127/38/29	332/255/154/117	133/128/154/117
	5	1328/508/152/116	330/128/40/30	332/254/152/116	133/128/154/116
sub		c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5
	1,2,3,4	92/204/461/739/610	23/51/115/185/152	92/204/461/370/129	92/102/116/124/122
	5	92/204/460/740/608	23/51/116/184/154	92/204/460/370/132	92/102/115/124/122
text		c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5
	1,2,3	135/78/122/269/1502	34/20/30/67/375	135/78/122/269/376	135/78/122/135/126
	4	135/79/121/269/1501	34/19/31/67/376	135/79/121/269/376	135/79/121/135/126
	5	136/79/121/268/1501	33/19/31/68/376	136/79/121/268/376	136/79/121/134/126
mar		c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5
	1,2,3,4	83/171/320/683/848	21/43/80/171/212	83/171/320/171/424	83/171/107/137/122
	5	84/172/320/684/848	20/42/80/170/212	84/172/320/171/424	84/172/107/137/122
mal		c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5	c1/c2/c3/c4/c5
	1,2	227/480/992/312/95	56/120/248/78/24	227/480/331/312/95	114/120/124/104/95
	3,4	226/480/992/312/95	57/120/248/78/24	226/480/331/312/95	113/120/124/104/95
	5	226/480/992/312/96	57/120/248/78/23	226/480/331/312/96	113/120/124/104/96

number of the training epochs are empirically determined for the fine-grained classification.

III. EXPERIMENTS AND RESULTS

In this study, the fine-grained subordinate classification is performed for the 7 different semantic attributes of lung nodules in the LIDC CT images. To illustrate the efficacy of the data over-sampling for the minority classes with the WGAN, five schemes are implemented for comparison. The first scheme, denoted as ORI, performs no over-sampling and data augmentation on the training data, whereas the second scheme, named AUG, carries out the standard data augmentation adopted in [7] and [19] on the training data, where each image can be rotated randomly with the degree in the range from 0 to 359. Afterward, the image may be flipped horizontally or vertically with probability of 0.5. The rest three schemes are implemented with the GAN-based synthetic over-sampling technique which adopt the typical GAN [24], DCGAN [25] and WGAN [26] respectively for data synthesis. All five schemes use the same CNNs with the architecture shown in Table 2 for the purpose of the fine-grained classification.

Since every CT scan in the LIDC dataset was read by 4 radiologists, each nodule can be possibly annotated by at least one radiologist. For nodules annotated by at least 2 radiologists, the annotated scores of the semantic attributes from different radiologists are averaged as the ground truth labels [10], [11], [20], [21], [41].

A. DATA OVER-SAMPLING AND EXPERIMENTAL SETTINGS

In this study, the data over-sampling for the minority classes, including the AUG, GAN, DCGAN and WGAN schemes, is evaluated with the 5-fold cross validation (CV). For each fold in the CV, the over-sampling is performed on the training dataset, whereas the validation data remains unchanged for the latter fine-grained classification. For fair comparison, the numbers of the synthesized samples of each class in each fold in the AUG, GAN, DCGAN and WGAN schemes are the same. Meanwhile, the data partitions in the 5-fold CV are the same for the ORI, GAN, DCGAN and WGAN schemes. Table 3 shows the details in the 5-fold CV data partitions for the five schemes and the batch size configuration in training GAN, DCGAN and WGAN. The batch size in training the WGAN, is mostly set as the same as the size of the training samples to obtain synthetic samples as many as possible. For the special cases like class 2 in ''lob'' and class 5 in ''sub'', the batch size are cut down for better training of the discriminator. However, the batch size setting in training the WGAN can not perform well in training the GAN and the DCGAN in practice. Therefore, we set the batch size a bit smaller in training the GAN and the DCGAN. Specifically, the batch size settings can be also found in Table 3.

The training of the GANs is conducted on 32-core Intel Xeon CPU E5-2620 and 128GB memory machine equipped with an NVIDIA Tesla M40 (24GB on-board memory) GPU card. The number of the synthesized samples for a minority class is determined by the size difference between the majority and the minority classes to make the samples of the minority and the majority classes as equal as possible. For each minority class, the training iterations is empirically set as 20000 to obtain the synthesized samples with quality. Fig. 6 illustrates the synthetic quality w.r.t. the number of generator iterations. As can be observed in Fig. 6, the synthetic quality can be reasonably good in the iteration of 20000. The training process of the WGAN is much more stable than the GAN and DCGAN.

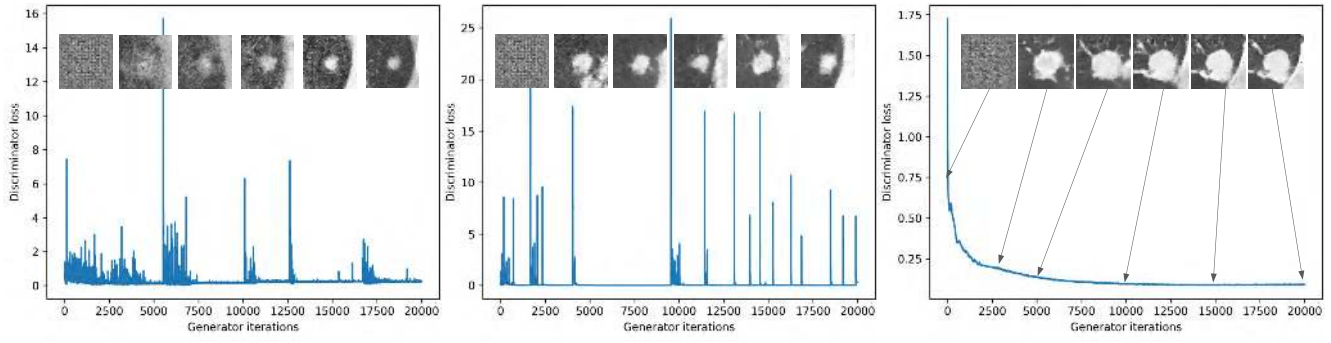


FIGURE 6. The training process and the synthetic quality with the discriminator loss for the GAN, DCGAN and WGAN, respectively.

B. FINE-GRAINED CLASSIFICATION FOR SEMANTIC ATTRIBUTES

The normal CNNs with architecture described in Table 2 are used for the task of the fine-grained classification. Specifically, the 7 classification CNNs are trained for the 7 semantic attributes of lung nodules. The fine-grained CNN classifiers are also evaluated with the 5-fold CV that share the same data partitions in the 5-fold CV of the over-sampling schemes in Table 3.

In this study, three assessment metrics are adopted to profile the performance of the subordinate classification and indirectly illustrate the efficacy of the over-sampling schemes. The first and the second assessment metrics are the F_1 score and the extended G-mean, which are derived by the basic metrics of the precision and recall. The third metric is absolute distance defined in e.q. (9) for better evaluating the classification performance of the subordinate classes that share very similar properties of the same base attribute class [20], [21].

The F_1 score is a standard accuracy of a multi-class classification problem by considering both precision and recall. Specifically, the F_1 score for the class i can be computed as

$$F_1(i) = \frac{2 * R(i) * P(i)}{R(i) + P(i)}, \tag{7}$$

where $P(i)$ and $R(i)$ are the precision and recall of the class i, respectively. Accordingly, the overall F_1 score for one semantic attribute can be simply obtained by averaging the F_1 scores of all k subordinate classes as $\sum_{i=1}^{k} F_1(i)/k$. The second extended G-mean metric [38], [42] is the geometric mean of the recalls over all k subordinate classes for one semantic attribute. The extended G-mean can reflect overall sensitivity for one semantic attribute. Specifically, the extended G-mean, $\overline{G}$, over all k subordinate classes for one attribute can be defined as

$$\overline{G} = (\prod_{i=1}^{k} R(i))^{\frac{1}{k}}, \tag{8}$$

where $R(i)$ is the recall of the class i. In general, larger values of F_1 score and G-mean suggest the better agreement between the predicted results and the labeled ground truths.

On the other hand, because the subordinate classes of each semantic attribute share very similar properties of the same base attibute class, the distance between the labeled class and the predicted class can be a referential index to reflect the relations of the subordinate classes [20], [21]. Accordingly, the metric of absolute distance is adopted to illustrate how close the prediction results to the true labeled class i, denoted as $d(i)$, which can be calculated as

$$d(i) = \frac{1}{N_i} \sum_{n=1}^{N_i} |v_n - \widetilde{v_n}|, \tag{9}$$

where N_i is the total number of the samples for the class i, and v_n and $\widetilde{v_n}$ are the true label, i.e., i, and predicted label of the sample n, respectively. Different to the metrics of F_1 score and G-mean, smaller absolute distance values suggest better performance of the classification method.

C. PERFORMANCE COMPARISON

To illustrate the efficacy of the data over-sampling techniques for the data imbalance issue, five schemes of ORI, AUG, GAN, DCGAN and WGAN are implemented for the fine-grained classification of the seven different semantic attributes of the lung nodules. Table 4 reports the overall performances in terms of the metrics of F_1 score, G-mean and absolute distance for the seven semantic attributes w.r.t. the five schemes. Since all five schemes are evaluated with the 5-fold CV, the mean $\pm$ standard deviation statistics for the three assessment metrics over the 5 folds are reported in Table 4.

As can be observed in Table 4, both the GAN and DCGAN schemes can achieve relatively higher F_1 scores and G-means and less absolute distances for most semantic attributes by comparing to the ORI scheme. Meanwhile, the DCGAN scheme performs a slight better than the GAN scheme. This may suggest that the synthetic data augmentation can slightly address the data imbalance issue for the CNN fine-grained classification. In addition, the AUG scheme can achieve slightly higher F_1 scores and G-means and less absolute distances for most semantic attributes by comparing to the

TABLE 4. Performance summary in terms of F_1 score, G-mean and Absolute distance for the 7 semantic attributes w.r.t. the five schemes over all five folds in CV.

Metric	Scheme	lob	sphe	spic	sub	text	mar	mal
F_1 score	ORI	0.25±0.03	0.24±0.01	0.23±0.01	0.29±0.02	0.26±0.03	0.30±0.03	0.37±0.04
	GAN	0.27±0.03	0.26±0.03	0.24±0.02	0.30±0.02	0.28±0.03	0.32±0.04	0.39±0.02
	DCGAN	0.27±0.02	0.27±0.02	0.28±0.03	0.32±0.02	0.31±0.02	0.32±0.02	0.40±0.04
	AUG	0.28±0.03	0.26±0.04	0.29±0.01	0.34±0.00	0.33±0.04	0.33±0.01	0.42±0.04
	WGAN	**0.35±0.03**	**0.32±0.02**	**0.33±0.02**	**0.36±0.01**	**0.40±0.02**	**0.36±0.03**	**0.46±0.03**
G-mean	ORI	0.09±0.11	0.00±0.00	0.13±0.07	0.13±0.11	0.08±0.10	0.26±0.04	0.33±0.05
	GAN	0.19±0.11	0.13±0.12	0.18±0.03	0.21±0.11	0.17±0.09	0.28±0.05	0.35±0.02
	DCGAN	0.18±0.11	0.13±0.13	0.22±0.03	0.27±0.03	0.23±0.02	0.29±0.03	0.36±0.06
	AUG	0.19±0.10	0.10±0.12	0.25±0.02	0.30±0.00	0.25±0.05	0.31±0.02	0.38±0.05
	WGAN	**0.33±0.04**	**0.30±0.03**	**0.30±0.02**	**0.34±0.01**	**0.37±0.03**	**0.35±0.04**	**0.44±0.04**
Absolute distance	ORI	1.11±0.04	0.97±0.02	1.20±0.04	1.03±0.04	1.18±0.10	1.05±0.06	0.81±0.06
	GAN	1.10±0.03	0.98±0.02	1.17±0.03	1.06±0.03	1.14±0.08	1.01±0.04	0.79±0.03
	DCGAN	1.08±0.04	0.99±0.04	1.09±0.05	1.02±0.07	1.09±0.07	1.00±0.03	0.77±0.05
	AUG	1.06±0.05	0.97±0.04	1.11±0.01	0.99±0.02	0.99±0.07	0.95±0.03	0.75±0.05
	WGAN	**0.96±0.05**	**0.93±0.04**	**1.03±0.02**	**0.96±0.02**	**0.85±0.05**	**0.91±0.03**	**0.72±0.05**

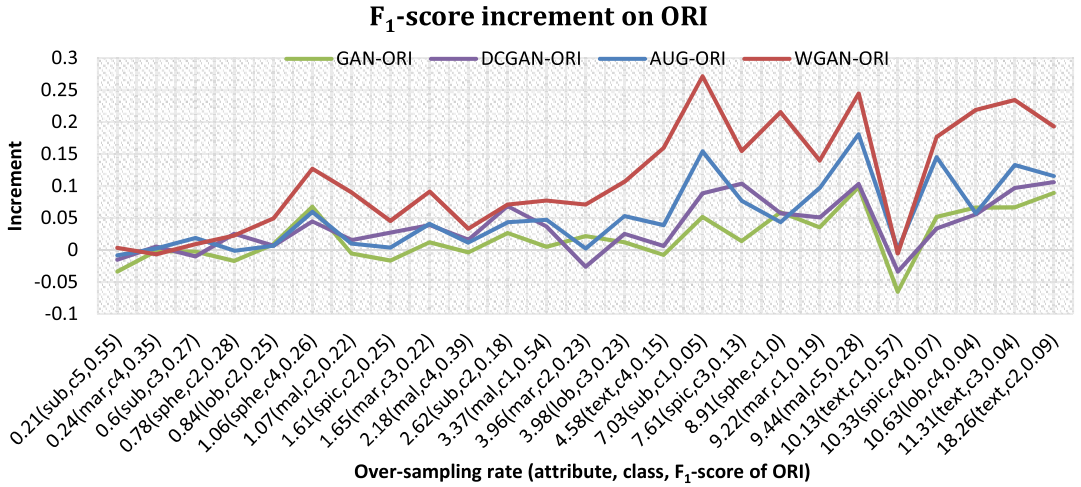


FIGURE 7. Performance boosting analysis of F_1 score w.r.t. the over-sampling rate for the GAN, DCGAN, AUG and WGAN schemes to the baseline scheme ORI.

DCGAN scheme. This indicates the efficacy of the conventional data augmentation method commonly used in many deep learning works. Besides, it is worth noting that the G-mean performance in the "sphe" attribute of the ORI scheme is 0. It is because the recall of its subordinate class 1 is 0 and hence the overall G-mean is resulted as 0 due to the operation of geometric mean. The sample number of the subordinate class 1 is 124, which is a minority class. Accordingly, it can be suggested that data imbalance is a critical impact factor for fine-grained classification.

On the other hand, it can be also found in Table 4 that the WGAN scheme can perform much better than the AUG for all semantic attributes tasks. It may be because that the WGAN-based synthetic data augmentation method attempts to approximate the underlying data distributions of the subordinate classes in each attribute. The conventional data augmentation (AUG) method doesn't consider the underlying data distribution and thus the boosting of performance may be limited. The reasons of that the performances of the GAN and DCGAN schemes do not provide better help might be the not enough stable training process of the GAN and DCGAN, particularly in the small samples situation. Therefore, the synthesized data from the GAN and DCGAN may sometimes not very representative for the subordinate classes comparing to the WGAN.

D. PERFORMANCE BOOSTING WITH OVER-SAMPLING SCHEMES

To further illustrate the efficacy of the over-sampling technique on the performance trends of all minority classes, we carry out the performance boosting analysis w.r.t. the factor of the over-sampling rate of all minority classes. The over-sampling rate is defined as $\frac{num(major)-num(minor)}{num(minor)}$, where $num(major)$ means the sample number of a majority class and $num(minor)$ means the sample number of a minority class. The operation of the over-sampling makes the sample number of the minority classes are equal to that of the majority class. The performance boosting here is defined as the difference between the over-sampling schemes of the GAN, DCGAN, AUG or WGAN to the baseline scheme ORI in terms of either F_1 score or absolute distance. Meanwhile, it is worth noting

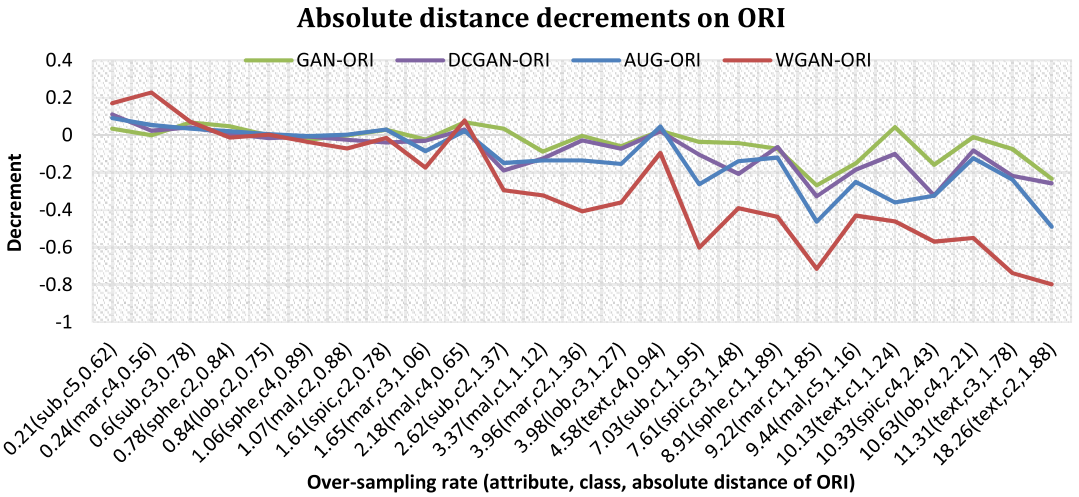


FIGURE 8. Performance boosting analysis of absolute distance w.r.t. the over-sampling rate for the GAN, DCGAN, AUG and WGAN schemes to the baseline scheme ORI.

again that the four over-sampling schemes of GAN, DCGAN, AUG and WGAN are only performed in the training phase data of the CNN fine-grained classification.

Fig. 7 illustrates the performance boosting in terms of F_1 score for the GAN, DCGAN, AUG and WGAN schemes over the ORI scheme w.r.t. all minority subordinate classes of the seven semantic attributes. The horizontal axis in Fig. 7 is sorted with the over-sampling rate of each subordinate class in ascending order, whereas each index in the horizontal axis is supplemented with 3-tuple of the corresponding semantic attribute, subordinate class and the F_1 score of ORI. As can be observed in Fig. 7, the performance boosting on both AUG and WGAN schemes can be mostly positive. The WGAN scheme can gain larger performance boosting than the other three schemes, particularly for the cases with larger over-sampling rates, i.e., the subordinate classes with samples significantly less than its majority class. It is worth noting that the DCGAN scheme is comparable to the AUG scheme when the over-sampling rate is smaller than 3.37, whereas the DCGAN scheme doesn't provide better help than the AUG scheme when the over-sampling rate is larger than 3.96. This suggests that the fact of the small original training samples might be an unfavorable condition for the DCGAN. In contrast, the WGAN scheme performs still efficiently and robustly even if the original training samples are very small.

Fig. 8 illustrates similar performance boosting analysis to Fig. 7 in terms of the absolute distance metric. Different to the F_1 score, the larger negative difference of the absolute distance suggests better performance boosting. It can be found that the better performance can be achieved when the over-sampling rate is larger than 2.5. Meanwhile, the WGAN scheme can also outperform the GAN, DCGAN and AUG schemes with the metric of absolute distance.

It can be observed in Fig. 7 that the over-sampling rate index of 10.13 (text, c1, 0.57) suggests almost no performance boosting for either AUG or WGAN.

TABLE 5. The classification ratios of the subordinate class1 over all subordinate classes of the "text" attribute w.r.t the ORI, GAN, DCGAN, AUG and WGAN schemes. For example, the notation of "c1→c3" means the proportion of true samples of the subordinate class1 being classified as the subordinate class3.

text	c1→c1	c1→c2	c1→c3	c1→c4	c1→c5
ORI	0.61	0.04	0.06	0.09	0.20
GAN	0.48	0.20	0.07	0.08	0.18
DCGAN	0.55	0.12	0.12	0.07	0.14
AUG	0.65	0.12	0.07	0.01	0.15
WGAN	0.57	0.21	0.12	0.06	0.04

However, the same index in Fig. 8 illustrates the performance boosting with the AUG and WGAN schemes. It is because that the F_1 score metric can only reflect the miss-classification but not illustrate the erroneous degree. Since there exists class relation among the consecutive subordinate classes in all semantic attributes, the erroneous degree shall suggest that the error of miss-classifying subordinate class 1 into class 5 can be larger than the error of miss-classifying class 1 into class 2. The erroneous degree can be reflected with the absolute distance metric. To further investigate the underlying cause of performance discordance for the index 10.13 in Figs. 7 and 8, we report the classification ratios of the subordinate class 1 (c1) into all subordinate classes (c1,c2,c3,c4,c5) in the attribute "text" w.r.t. the five schemes of the ORI, GAN, DCGAN, AUG and WGAN in Table 5. As can be found, the miss-classification ratios of "c1→c5" in the ORI and AUG schemes are perceivably larger than the ratio in the WGAN scheme, whereas the situation is just opposite for the ratios of "c1→c2". Unfortunately, the miss-classifications of the cases "c1→c2" and "c1→c5" are treated equally for the computation of F_1 score. The absolute distance on the other hand can reflect the difference between the cases "c1→c2" and "c1→c5" with erroneous degree. Accordingly, the performance discordance in Figs. 7 and 8 for the index 10.13 can be explained.

TABLE 6. Performance summary in terms of F_1 score, G-mean and Absolute distance for the 7 semantic attributes w.r.t. the schemes of the GAN, DCGAN, AUG and WGAN over all five folds in CV when more realistic images are augmented by comparing to Table 4.

Metric	Scheme	lob	sphe	spic	sub	text	mar	mal
F_1 score	GAN	0.29±0.01,↑0.02	0.27±0.03,↑0.01	0.30±0.03,↑0.06	0.29±0.04,↓0.01	0.25±0.03,↓0.03	0.25±0.03,↓0.07	0.41±0.02,↑0.02
	DCGAN	0.36±0.01,↑0.09	0.29±0.03,↑0.02	0.31±0.04,↑0.03	0.34±0.03,↑0.02	0.27±0.04,↓0.04	0.31±0.02,↓0.01	0.42±0.05,↑0.02
	AUG	0.34±0.03,↑0.06	0.30±0.02,↑0.04	0.34±0.01,↑0.05	0.36±0.03,↑0.02	0.36±0.05,↑0.03	0.32±0.03,↓0.01	0.47±0.03,↑0.05
	WGAN	**0.39±0.02**,↑0.04	**0.36±0.02**,↑0.04	**0.39±0.01**,↑0.06	**0.39±0.00**,↑0.03	**0.42±0.04**,↑0.02	**0.40±0.01**,↑0.04	**0.52±0.01**,↑0.06
G-mean	GAN	0.27±0.01,↑0.08	0.15±0.12,↑0.02	0.27±0.04,↑0.09	0.24±0.06,↑0.03	0.18±0.03,↑0.01	0.19±0.02,↓0.09	0.36±0.03,↑0.01
	DCGAN	0.33±0.02,↑0.15	0.15±0.12,↑0.02	0.26±0.05,↑0.04	0.30±0.03,↑0.03	0.15±0.12,↓0.08	0.28±0.02,↓0.01	0.38±0.06,↑0.02
	AUG	0.29±0.03,↑0.10	0.26±0.02,↑0.16	0.30±0.01,↑0.05	0.20±0.11,↓0.10	0.28±0.05,↑0.03	0.22±0.03,↓0.09	0.39±0.03,↑0.01
	WGAN	**0.38±0.03**,↑0.05	**0.35±0.02**,↑0.05	**0.37±0.02**,↑0.07	**0.37±0.01**,↑0.03	**0.40±0.04**,↑0.03	**0.39±0.01**,↑0.04	**0.51±0.01**,↑0.07
Absolute distance	GAN	1.08±0.04,↓0.02	0.93±0.02,↓0.05	1.11±0.03,↓0.06	1.13±0.08,↑0.07	1.32±0.15,↑0.18	1.28±0.05,↑0.27	0.73±0.05,↓0.06
	DCGAN	0.97±0.07,↓0.11	0.93±0.04,↓0.06	1.03±0.06,↓0.06	1.02±0.07,↓0.00	1.23±0.10,↑0.14	1.08±0.04,↑0.08	0.77±0.07,↓0.00
	AUG	0.98±0.04,↓0.08	0.91±0.05,↓0.06	0.97±0.03,↓0.14	0.95±0.07,↓0.04	0.91±0.09,↓0.08	1.05±0.08,↑0.10	0.69±0.05,↓0.06
	WGAN	**0.88±0.05**,↓0.08	**0.87±0.02**,↓0.06	**0.91±0.04**,↓0.12	**0.90±0.02**,↓0.06	**0.80±0.04**,↓0.05	**0.87±0.03**,↓0.04	**0.65±0.03**,↓0.07

E. PERFORMANCE IMPROVEMENT WITH MORE REALISTIC AUGMENTED DATA

To illustrate the efficacy of the data over-sampling techniques for the data imbalance issue when more realistic samples are augmented, we attempt to augment more data for each subordinate class in all semantic attributes. With the operation of the more data augmentation, the sample number of each subordinate class has been **doubled** on the basis of the schemes of the GAN, DCGAN, AUG and WGAN in Table 4. Table 6 reports the overall performances in terms of the metrics of F_1 score, G-mean and absolute distance for the seven semantic attributes w.r.t. the schemes of the GAN, DCGAN, AUG and WGAN. Meanwhile, the performance of increment (↑) and decrement (↓) comparing to Table 4 w.r.t the F_1 score, G-mean and absolute distance are also supplemented in Table 6.

As can be observed in Table 6, the performance of the fine-grained classification of the WGAN scheme has been improved for all semantic attribute tasks when double amounts of data for each subordinate class in all semantic attributes are synthesized. The WGAN scheme still performs much better than the schemes of the GAN, DCGAN and AUG. The performance of the schemes of the GAN, DCGAN and AUG has also been boosted for most semantic attributes with more augmented data. However, there are some degrees of performance degradation in the schemes of the GAN, DCGAN and AUG for the semantic attributes ''sub'', ''text'' and ''mar''. This may suggest that the WGAN scheme can be more effective and robust than the schemes of the GAN, DCGAN and AUG for addressing the data imbalance problem.

IV. DISCUSSION AND CONCLUSION

In this paper, the WGAN technique is exploited to address the data imbalance issue which commonly exists in the medical image classification problems. The WGAN technique is able to estimate the underlying distribution of a minority class domain and hence can synthesize plausible samples to mitigate the data imbalance issue for the performance boosting. The WGAN-based data synthetic over-sampling technique is specifically applied for the fine-grained classification of the 7 nodule semantic attributes in the public LIDC dataset. As can be found in Fig. 1, the sample distribution of the subordinate classes are very imbalanced, the optimization of classification or regression [20], [21] can easily favor the majority classes. In such case, the accuracy over the minority classes can be scarified to attain smaller classification or regression errors. To clearly illustrate the effectiveness of WGAN-based data synthetic over-sampling technique for the data imbalance issue, a CNN architecture is employed for the fine-grained classification of the 7 semantic attributes. Referring to Tables 4, 6 and Figs. 7, 8, the experimental results suggest the efficacy of the WGAN scheme for the performance boosting on the fine-grained classification of the 7 semantic attributes, particularly for those minority subordinate classes.

Meanwhile, it is also shown that the WGAN-based data synthetic over-sampling technique can be more effective than the conventional data augmentation (AUG) scheme, which is commonly practiced in many deep learning works. It is suggested that more helpful synthesized samples can be obtained by considering the underlying distributions of the minority classes. On the other hand, the synthesized data from the schemes of the GAN and DCGAN do not provide better help in the data imbalance problem than the synthesized data from the AUG scheme. Referring to Fig. 6, the training processes of the generators in the GAN and DCGAN can be very unstable in terms of the discriminator loss. The synthesized data from the GAN and DCGAN may sometimes turn out to be implausible or not very representative for the class of minority. Therefore, helpful synthesized data may not be easily obtained with a systematic tuning for the training of the GAN and DCGAN. By contrast, the training of the generator in the WGAN scheme is relatively stable and therefore the quality of the synthesized data from the WGAN scheme can be better assured.

In the medical context, the minority classes can be very important. Since for some subtypes or stages of one disease can be very difficult to collect, it will lead to significant data imbalance situation, and then render the classification problem very arduous. For example, the subordinate class 3 in the attribute ''text'' refers to the sub-solid nodules, which are relatively rare in the LIDC data but important as this type of nodules are highly associated to the subtype of adenocarcinomas. Therefore, accurate identification of sub-solid nodules can be helpful for the determination of adenocarcinomas for

more precise diagnosis and treatment recommendation in the computer-aided diagnosis (CAD) application. Without the support of sufficient training data, promising performance of the fine-grained classification may not be easily attained. In such case, the tackle of the data imbalance matters for many AI applications in medicine.

Important as it is, the data imbalance issue has been less explicitly explored in previous computer-aided medical image analysis studies. In particular, with the popularization of data-hungry deep learning paradigms, the data imbalance issue may be getting important. In this study, the efficacy of the WGAN-based over-sampling scheme is demonstrated to address the serious data imbalance issue for the fine-grained classification on the LIDC dataset. The fine-grained classification is per se a quite difficult problem and satisfactory performance can't not be easily achieved [43], [44]. For the fine-grained classification of the nodule semantic attributes, the boundary between the adjacent subordinate classes can be very ambiguous and subjective. The difficulty of this problem is further exacerbated by the severe data imbalance issue and turns out to be very arduous. Although promising regression performance was reported in [20] and [21], low regression values don't necessarily reflect high prediction accuracy for the minority classes. As the optimization may favor the majority classes and scarify the minority classes, the data imbalance issue shall be explicitly addressed, but unfortunately was not tackled in many previous studies [20], [21] on the LIDC dataset. The WGAN-based approach presented in this paper is shown to be better than the conventional data augmentation technique used in many previous deep learning works and the fine-grained classification without any data augmentation. Accordingly, it may thus shed a light on using the WGAN technique for data augmentation for the future imbalanced deep learning studies.

The major limitation of this study lies in that we perform minority over-sampling on individual semantic attributes, and hence the whole over-sampling process can be little bit tedious. The future work will explore on the relations among the semantic attributes for the data synthesis to simplify the over-sampling process as well as to boost the synthetic quality even further.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[2] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*. Cham, Switzerland: Springer, 2016, pp. 424–432.

[3] D. C. Cirean, G. Alessandro, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 2852–2860.

[4] C. Bian, R. Lee, Y.-H. Chou, and J.-Z. Cheng, "Boundary regularized convolutional neural network for layer parsing of breast anatomy in automated whole breast ultrasound," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 259–266. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-66179-7_30

[5] B. Lei *et al.*, "Segmentation of breast anatomy for automated whole breast ultrasound images with boundary regularized convolutional encoder–decoder network," *Neurocomputing*, vol. 321, pp. 178–186, Dec. 2018.

[6] H. R. Roth *et al.*, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1170–1181, May 2016.

[7] H. R. Roth *et al.*, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2014, pp. 520–527.

[8] A. A. A. Setio *et al.*, "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[9] J. Z. Cheng *et al.*, "Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans," *Sci. Rep.*, vol. 6, p. 24454, Apr. 2016.

[10] W. Shen *et al.*, "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification," *Pattern Recognit.*, vol. 61, pp. 663–673, Jan. 2017.

[11] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, Eds. Cham, Switzerland: Springer, 2015, pp. 588–599.

[12] X. Tu *et al.*, "Automatic categorization and scoring of solid, part-solid and non-solid pulmonary nodules in ct images with convolutional neural network," *Sci. Rep.*, vol. 7, no. 1, p. 8533, Sep. 2017.

[13] Q. Wang *et al.*, "Low-shot multi-label incremental learning for thoracic diseases diagnosis," in *Neural Information Processing*, L. Cheng, A. C. S. Leung, and S. Ozawa, Eds. Cham, Switzerland: Springer, 2018, pp. 420–432.

[14] L. Wu, J. Cheng, S. Li, B. Lei, T. Wang, and D. Ni, "FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1336–1349, May 2017.

[15] H. Chen *et al.*, "Ultrasound standard plane detection using a composite neural network framework," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1576–1586, Jun. 2017.

[16] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[17] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[18] X. Zhang, H. Su, L. Yang, and S. Zhang, "Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5361–5368.

[19] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[20] S. Chen, D. Ni, J. Qin, B. Lei, T. Wang, and J.-Z. Cheng, "Bridging computational features toward multiple semantic features with multi-task regression: A study of ct pulmonary nodules," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham, Switzerland: Springer, 2016, pp. 53–60.

[21] S. Chen *et al.*, "Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in CT images," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 802–814, Mar. 2017.

[22] S. G. Armato *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on ct scans," *Acad. Radiol.*, vol. 14, no. 12, pp. 1455–1463, 2011.

[23] M. F. McNitt-Gray *et al.*, "The lung image database consortium (LIDC) data collection process for nodule detection and annotation," *Acad. Radiol.*, vol. 14, no. 12, pp. 1464–1474, 2007.

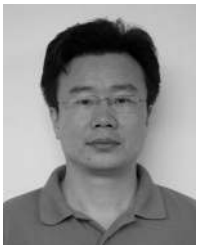
[24] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Nov. 2016, pp. 1–16.

[26] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 214–223.

[27] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn. (ICML JMLR)*, vol. 48, 2016, pp. 1060–1069.
[28] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5908–5916.
[29] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Jul. 2017, pp. 5967–5976.
[30] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
[31] D. Nie *et al.*, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 12, pp. 2720–2730, Dec. 2018.
[32] Y. Zhang, J.-Z. Cheng, L. Xiang, P.-T. Yap, and D. Shen, "Dual-domain cascaded regression for synthesizing 7T from 3T MRI," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 410–417.
[33] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose CT," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017.
[34] W. Dai *et al.* (Apr. 2017). "SCAN: Structure correcting adversarial network for organ segmentation in chest X-rays." pp. 1–10. [Online]. Available: https://arxiv.org/abs/1703.08770v1
[35] Z. Li, Y. Wang, and J. Yu, "Reconstruction of thin-slice medical images using generative adversarial network," in *Machine Learning in Medical Imaging*, Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, Eds. Cham, Switzerland: Springer, 2017, pp. 325–333.
[36] A. P. Reeves and A. M. Biancardi. (Oct. 27, 2011). *The Lung Image Database Consortium (LIDC) Nodule Size Report*. [Online]. Available: http://www.via.cornell.edu/lidc/
[37] A. P. Reeves *et al.*, "The lung image database consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements," *Acad. Radiol.*, vol. 14, no. 12, pp. 1475–1485, 2007.
[38] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 238–251, Jan. 2016.
[39] A. Orriols-Puig and E. Bernadó-Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Comput.*, vol. 13, no. 3, p. 213, May 2008.
[40] I. J. Goodfellow. (Apr. 2017). "NIPS 2016 tutorial: Generative adversarial networks," pp. 1–57. [Online]. Available: https://arxiv.org/abs/1701.00160
[41] F. Han *et al.*, "A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database," in *Proc. IEEE Int. Conf. Med. Imag. Phys. Eng. (ICMIPE)*, Oct. 2013, pp. 14–18.
[42] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalanced class distribution," in *Proc. 6th Int. Conf. Data Mining*, Dec. 2006, pp. 592–602.
[43] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 4, pp. 1119–1130, Aug. 2012.
[44] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3973–3981.

QINGFENG WANG received the M.S. degree in computer science from the Southwest University of Science and Technology, Mianyang, China, in 2014. She is currently pursuing the Ph.D. degree with the School of Software Engineering, University of Science and Technology of China, Heifei, China. Her current research interests include computer-aided diagnosis, medical image analysis, pattern recognition, machine learning, and class imbalance learning.



XUEHAI ZHOU is currently the Executive Dean of the School of Software Engineering, University of Science and Technology of China, and a Professor with the School of Computer Science. His research interests include the areas of software engineering, operating systems, and distributed computing systems. He serves as the General Secretary for the Steering Committee on Computer College Fundamental Lessons and the Technical Committee on Open Systems of the China Computer Federation.

CHAO WANG received the B.S. and Ph.D. degrees from the School of Computer Science, University of Science and Technology of China, Hefei, China, in 2006 and 2011, respectively, where he is currently an Associate Researcher with the School of Software Engineering. He has authored more than 60 publications and patents, including publications in IEEE TC, ACM TCBB, and TACO as well as through the FPGA conference. His research interests include multicore and reconfigurable computing. He is an Editorial Board Member of MICPRO and IET CDT and a Guest Editor of TCBB and IJPP.

ZHIQIN LIU received the M.S. degree in computer engineering from the University of Electronic Science and Technology, Chengdu, China, in 1994. She is currently a Professor with the School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China. Her current research interests include medical image analysis, computer-aided diagnosis, and high-performance computing.

JUN HUANG received the M.S. degree in computer science from the Southwest University of Science and Technology, Mianyang, China, in 2014. He is currently pursuing the Ph.D. degree with the Graduate School, China Academy of Engineering Physics, Mianyang. He is currently a Teaching Assistant with the School of Computer Science and Technology, Southwest University of Science and Technology. His current research interests include big data, scientific computing, and high-performance computing.

YING ZHOU received the M.D. degree from Southern Medical University, Guangzhou, China, in 2018. She is currently an Attending Doctor with the Radiology Department, Mianyang Central Hospital, Mianyang, China. Her current research interests include texture analysis of tumor image and machine learning.

CHANGLONG LI received the B.S. and Ph.D. degrees from the Department of Computer Science, University of Science and Technology of China, Hefei, China, in 2012 and 2018, respectively. He has authored or co-authored more than 20 publications, technical reports, and patents. His research interests include big data and mobile cloud computing.

HANG ZHUANG received the B.S. and Ph.D. degrees from the Department of Computer Science, University of Science and Technology of China, Hefei, China, in 2012 and 2018, respectively. His research interests include cloud computing, machine learning, parallel and distributed computing, and natural language processing.

JIE-ZHI CHENG received the Ph.D. degree in biomedical engineering from National Taiwan University, Taipei, Taiwan, in 2013. He is currently a Vice President of research and development with Shanghai United Imaging Intelligence, Shanghai, China. His current research interests include medical image analysis, computer-aided diagnosis and intervention, pattern recognition, and machine learning. He has co-authored more than 40 scientific articles in several medical image analysis journals of IEEE TMI, *Radiology*, *Medical Physics*, and *Ultrasound in Medicine and Biology* and first tier medical image conferences like IPMI and MICCAI.