symmetry

MDPI

Article

Synthetic Medical Images Using F&BGAN for Improved Lung Nodules Classification by Multi-Scale VGG16

Defang Zhao [1], Dandan Zhu [1], Jianwei Lu [1,2,*], Ye Luo [1,*] and Guokai Zhang [1]

[1] School of Software Engineering, Tongji University, Shanghai 201804, China; 1731534@tongji.edu.cn (D.Z.); dandanzhu@tongji.edu.cn (D.Z.); zhang.guokai@tongji.edu.cn (G.Z.)
[2] Institute of Translational Medicine, Tongji University, Shanghai 201804, China
* Correspondence: jwlu33@tongji.edu.cn (J.L.); yeluo@tongji.edu.cn (Y.L.); Tel.: +86-21-6958-9585 (Y.L.)

Received: 28 September 2018; Accepted: 16 October 2018; Published: 17 October 2018

Abstract: Lung cancer is one of the highest causes of cancer-related death in both men and women. Therefore, various diagnostic methods for lung nodules classification have been proposed to implement the early detection. Due to the limited amount and diversity of samples, these methods encounter some bottlenecks. In this paper, we intend to develop a method to enlarge the dataset and enhance the performance of pulmonary nodules classification. We propose a data augmentation method based on generative adversarial network (GAN), called Forward and Backward GAN (F&BGAN), which can generate high-quality synthetic medical images. F&BGAN has two stages, Forward GAN (FGAN) generates diverse images, and Backward GAN (BGAN) is used to improve the quality of images. Besides, a hierarchical learning framework, multi-scale VGG16 (M-VGG16) network, is proposed to extract discriminative features from alternating stacked layers. The methodology was evaluated on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset, with the best accuracy of 95.24%, sensitivity of 98.67%, specificity of 92.47% and area under ROC curve (AUROC) of 0.980. Experimental results demonstrate the feasibility of F&BGAN in generating medical images and the effectiveness of M-VGG16 in classifying malignant and benign nodules.

Keywords: Computer Aided Diagnosis (CAD); Generative Adversarial Network (GAN); multi-scale; pulmonary nodules

1. Introduction

Lung cancer is the most common cause of cancer-related death in America, which accounts for 26% and 25% in men and women, respectively. The five-year relative survival rate of lung cancer is only 18%. By screening with computed tomography (CT), there is potential for lung cancer to be diagnosed at an earlier stage, which has been shown to reduce lung cancer mortality by up to 20% [1]. Small masses of tissues found in the lungs, called lung/pulmonary nodules, have the potential to become cancerous [2]. Thereby, the early diagnosis of lung nodules is important to increase the likelihood of survival rate.

The traditional diagnostic method is to manually analyze lung CT scans by trained radiologists, which may be error-prone and time-consuming. Therefore, computer-aided diagnosis (CAD) systems have been developed, which are based on numeric image features to implement automatic classification of lung nodules as either benign or malignant and help design follow-up treatment plans. In general, the lung cancer diagnosis CAD system involves four phases: image processing, extraction of region of interest (ROI), feature selection and classification. Among them, feature selection and classification are key steps to improve the sensitivity and accuracy of the entire system. The conventional CAD systems [3–6] are fundamentally based on complex pattern recognition, which highly rely on image

processing to capture reliable features. Despite the favorable performance of these CAD systems in lung nodules analysis, the extracted features tend to be subjective, which limits the performance of the model to a certain extent.

In recent years, motivated by the creditable performance of neural network in the fields of computer vision, applying the deep learning technique in medical image has become a main trend that shows promising results. Consequently, various CAD systems based on neural network have been proposed to implement the classification of lung nodules [7–15]. Compared to traditional CAD systems, neural network based systems can automatically extract high-level features from the original images by using different network structures. Although the network structure can strongly influence the performance of CAD systems, the amount of training data has the biggest impact. Due to limited amount and diversity of medical image, deep learning based CAD systems may encounter some bottlenecks, for example, overfitting. Moreover, in some methods [9,15], image segmentation is used to extract lung nodules. Automatic nodule segmentation may affect classification results since these methods depend on initialization. Working on these segmented nodules may produce inaccurate features that lead to incorrect results. Given the provide citation's limitation, we attempt to develop a data generation method to enrich lung nodule images and use the generated data for pulmonary nodule classification.

Conventional data augmentation methods are mostly based on geometric transformations, such as flip, distort, crop, rotate, zoom, and so on. Recently, Goodfellow proposed generative adversarial networks (GAN) [16] which can generate images. Therefore, some investigations propose to enlarge medical image dataset using GAN [17–19]. Unlike these works, this paper proposes a new data augmentation technique, Forward and Backward GAN (F&BGAN), based on deep convolutional GAN (DCGAN) [20], which contains Forward GAN and Backward GAN. In the original formulation, DCGAN lacks a way to inversely map the real image space to the latent space. We propose a bi-directional learning framework Forward GAN to fill the above-mentioned gaps. In Forward GAN, we add a new pipeline and propose a new objective for the generator. Those improvements solve the gradient vanishing problem and mode collapse problem to a certain extent. Besides, to further improve the image quality, we use Backward GAN to process the generated images of Forward GAN. In Backward GAN, the discriminator acts as an encoder, while the generator is a decoder. Firstly, the encoder extracts features of images, and then the decoder generates high-quality images using these features.

Our Forward GAN is mostly related to Adversarially Learned Inference (ALI) [21] and Bidirectional GANs (BiGAN) [22]. However, there are two main differences between ALI, BiGAN and our Forward GAN: (1) The network structure is different. The ALI and BiGAN consist of three networks: (1) discriminator D; (2) generator G, which maps the noise vectors from latent space to image space; and (3) encoder E, which maps from real image space to latent space. Unlike them, our Forward GAN only has two networks, discriminator D and generator G. Our network D is not only a discriminator D, but also acts as an encoder E to map the real image space to latent space. (2) The objective is different. Their objective is to match two joint distributions. In our Forward GAN, except for the original objective of DCGAN, we employ an additional objective, which results in a cycle constraint that enables better communication between the network D and network G.

We present a multi-scale model M-VGG16 that uses VGG16 [23] as the structural backbone. As shown in Figure 1, there two main types of pulmonary nodules. We can see that the diameter of the nodules varies greatly. Consequently, it is crucial to extract features from fine scales during diagnosis of diverse diameter nodules. However, with the same receptive field, VGG16's ability to tolerate scale variations is limited. This means that VGG16 tends to extract specific scale features and this might lose some important information. The loss of detailed information of nodules can be an obstacle to the classification of lung nodules, especially for small diameter nodules. Therefore, multi-scale is essential in our model. Many multi-scale libraries have been proposed, e.g. scale-invariant feature transform (SIFT) [24]. However, SIFT is mathematically complicated and computationally heavy. In addition,

SIFT and many other multi-scale libraries are patent protected. Consequently, we design Multi-Scale Blocks (MSBs) to tackle this problem. MSB contains a series of filters with different kernel size. Filters with different kernel sizes can extract discriminative scale features, and this is significant in dealing with various appearance nodules. Therefore, MSB is conducive to extract scale-relevant features. Besides, our classification is based on lung image patch, which removes the inaccuracy introduced by image segmentation.

The contribution of our method is two-fold: (1) For data augmentation, we propose a new method F&BGAN, which can generate diverse high-quality images. (2) For lung nodules classification, we propose a multi-scale VGG16 network to extract more features. Extensive experimental results on Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset [25] further validate the superiority of the proposed method on lung nodules classification.

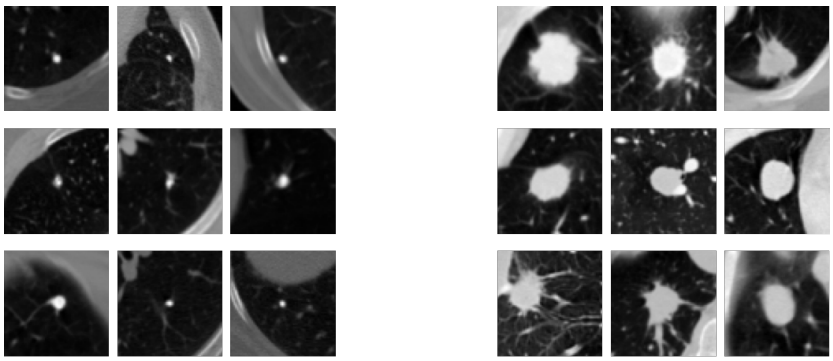

Figure 1. Different types of samples. Two main types of pulmonary nodules in LIDC-IDRI dataset. (**a**) small diameter nodules; (**b**) big diameter nodules.

The rest of the paper is organized as follows. Section 2 introduces the related works. Section 3 introduces the proposed methodology. Section 4 describes the experimental setup and analyses of the result. Section 5 discusses the impact and future of the project and concludes the presented research works.

2. Related Work

In the past, several methods have been proposed to classify lung nodules as benign or malignant using different algorithms. They can be generally divided into two categories: traditional methods and neural network based methods.

Traditional methods characterize the nodules with several quantitative features [26]. Some geometric feature descriptors, e.g., SIFT [24], local binary pattern (LBP) [27], etc., are combined with a traditional classification algorithm, e.g., support vector machines (SVM) [28], k-Nearest neighbor (KNN) [29], etc., to solve lung nodules classification problems, and this has achieved great success. For example, Farag et al. [3] used two geometric feature descriptors (SIFT and LBP) to extract features from the nodule candidates and used KNN to classify with an overall sensitivity of 86% and specificity of 97%. Orozco et al. [4] proposed a methodology that extracted eight texture features from the histogram and the gray level co-occurrence matrix, and then used SVM to classify lung tissues, with an accuracy rate of 82.66%. Krewer et al. [6] combined texture and shape features to classify malignant and benign lung nodules using several classifiers including Decision Trees (DT), KNN and SVM, with an accuracy rate of 90.91%. Parveen and Kavitha [5] proposed a method based on texture features using SVM, with a sensitivity rate of 91.38% and specificity rate of 89.56%.

Lately, neural network has become more popular in medical image diagnosis. Dandil et al. [9] used multiscale processing and Artificial Neural Networks (ANNs) to classify the nodules, which reached an accuracy of 90.63%, sensitivity of 92.30% and specificity of 89.47%. Hua et al. [8] proposed

a methodology using Deep Belief Network (DBN) and Convolutional Neural Network (CNN), with a sensitivity of 73.40% and specificity of 82.20% using DBN, and a sensitivity of 73.30% and specificity of 78.70% using CNN. Kumar et al. [10] proposed a method using Stacked Autoencoder (SAE), with an accuracy of 75.01%, sensitivity of 83.35%. Shen et al. [12] proposed a hierarchical learning framework—Multiscale Convolutional Neural Network (MCNN)—to capture nodule heterogeneity by extracting discriminative features from alternating stacked layers, with an accuracy of 86.84%. Cheng et al. [11] proposed a deep learning-based CADx system using stacked denoising autoencoder (SDAE), with the best accuracy of 95.6%, sensitivity of 92.4% and specificity of 98.9%. Shen et al. [7] presented a Multi-crop Convolutional Neural Network (MC-CNN) by using regions cropped from the convolutional feature maps which showed an accuracy of 87%, sensitivity of 77% and specificity of 93%. Kwajiri and Tezuka [13] used Convolutional Neural Network (CNN) and the Residual Network (ResNet) to classify the nodules, with a best sensitivity of 76.64% and specificity of 81.97% using CNN, and a best sensitivity of 89.50% and specificity of 89.38% using ResNet. Abbas [14] presented a multilayer combination of the convolutional neural network (CNN), recurrent neural networks (RNNs) and softmax linear classifies, with a sensitivity of 88% and specificity of 80%. The method proposed by da Silva et al. [15] used the evolutionary convolutional neural network to complete diagnosis, with the best sensitivity of 94.66%, specificity of 95.14%, and accuracy of 94.78%, as shown in Table 1

Table 1. An overview of representative related works on lung nodule classification.

Approach	Author	Year	Method
Traditional	Farag et al. [3]	2011	Texture features and KNN.
	Orozco et al. [4]	2013	Eight texture features and SVM.
	Krewer et al. [6]	2013	Texture and shape features with DT, KNN, and SVM.
	Parveen and Kavitha [5]	2014	Use GLCM to extract features and SVM to classify.
Neural network	Dandil et al. [9]	2014	ANNs for classification.
	Hua et al. [8]	2015	DBN and CNN for classification.
	Kumar et al. [10]	2015	SAE for classification.
	Shen et al. [12]	2015	Multi-scale CNN (MCNN) for classification.
	Cheng et al. [11]	2016	SDAE for classification.
	Shen et al. [7]	2017	Multi-crop CNN (MC-CNN) for classification.
	Kwajiri and Tezuka [13]	2017	CNN and ResNet to classify.
	Abbas [14]	2017	Integrate CNN and RNN for classification.
	da Silva et al. [15]	2017	Evolutionary CNN for classification.

Since Goodfellow proposed GAN [16], many GAN variants have been created. For example, Radford and Metz [20] proposed Deep Convolution GAN (DCGAN), which introduced a convolutional architecture. In addition to structure improvement and theoretical extension, another major innovative point is novel applications. Especially in the medical imaging domain, the main bottleneck is the data. Therefore, some recent works used GAN to cope with the small datasets and the limited number of annotated samples. Guibas et al. [18] proposed a two-stage GAN to generate retinal fundi images and vessel segmentation masks. Frid-Adar et al. [17] used GAN to enlarge the liver dataset. Chuquicusma et al. [19] used DCGAN to generate realistic lung nodules in the literature. Different from Chuquicusma's work, we generate data based on original nodule patches instead of segmented nodules. Moreover, we use the generated lung nodules patches to classify lung nodules for the first time.

3. Methodology

In this paper, we propose a model to train a generator to generate lung images firstly, and then train a classifier to distinguish between benign and malignant nodules. The overview of our model is shown in Figure 2.

Given a stack of lung CT images, we first extract the region of interest (ROI) containing the lung nodules. This will improve classification accuracy since the rest of lung tissue may affect the results. At the same time, this will reduce the training time because models only need to focus on smaller region. Therefore, we extract image patches containing lung nodules based on annotations and diagnostic information. The size of these image patches is 64 × 64. After that, we increase the amount of training data by F&BGAN, which includes two stages to generate high-quality synthetic images. Using original samples and generated samples, M-VGG16 captures discriminative features as fully as possible to classify nodules into benign and malignant.

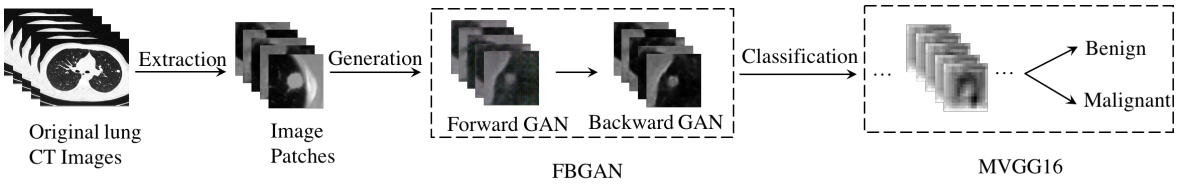

Figure 2. Overview of the whole structure. It is composed of a F&BGAN to generate diverse lung images (Section 3.1), and a multi-scale VGG16 network for classification (Section 3.2).

3.1. F&BGAN for Data Augmentation

Medical data are more difficult to obtain than natural data owing to privacy issues. As a result, many important tasks such as cancer classifications are hindered by lack of data. To tackle this problem, we propose a data augmentation method F&BGAN. As shown in Figure 3, F&BGAN consists of two parts: (1) Forward GAN, which is an image generator that can generate a bunch of diverse images; and (2) Backward GAN, which acts as a noise reducer and makes the generated images more realistic and high-quality. The two parts are seamlessly cascaded together, and the samples generated by Forward GAN are the input of Backward GAN.

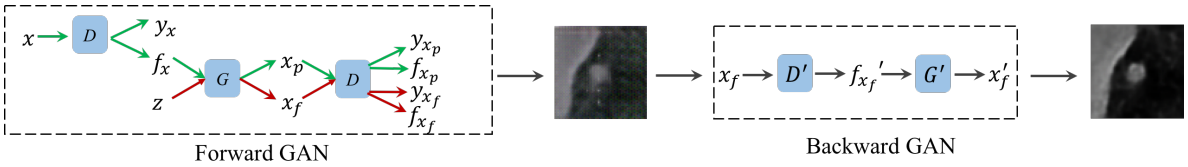

Figure 3. Illustration of the structure of F&BGAN, where x is the input image;, z is the input noise; D is a discriminator, which not only gives the classification result, such as y_x, y_{x_p} and y_{x_f}, but also gives the extracted features, such as f_x, f_{x_p} and f_{x_f}; G is a generative network; x_p, x_f are generated images; $D^{'}$ is a discriminator, which extracts features $f^{'}_{x_f}$ of the input; and $G^{'}$ is a generator that generates the final synthetic image $x^{'}_f$.

The structure of DCGAN is shown in Figure 4. Our model F&BGAN makes some improvements on the basis of DCGAN to get better performance. For the Forward GAN, the red pipeline of FGAN is the same as that of DCGAN, and a green pipeline is newly added. In the green pipeline, we use the network D to extract features f_x from the real images. Network G takes these features as input and learns to generate images x_p. By adding the input of real images, the network G can learn the basic characteristics of lung nodule images. This is beneficial to generate more realistic and diverse images and avoid mode collapse problem to a certain extent.

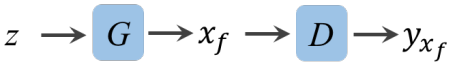

Figure 4. Illustration of the structure of DCGAN, where z is the input noise, G is a generative network, x_f are generated images, and D is a discriminator, which gives the classification result y_{x_f}.

Backward GAN is used to eliminate noises of the images generated by Forward GAN. By comparing Figure 3 with Figure 4, we can notice that the network structure of BGAN is inverse to that of DCGAN. In the network $D^{'}$ of BGAN, the generated images pass through several convolutional layers, which can extract features of the images and remove unwanted noises. Then, the network $G^{'}$ utilizes these features to generate new clear images.

3.1.1. Forward GAN.

As shown in Figures 5 and 6, the structures of generator (G) and discriminator (D) in Forward GAN are almost the same as that in DCGAN. The only difference is that we make a change in network D. In addition to the original Fully Connected (FC) layer used to discriminate, we add another FC layer to obtain the features of the input image. As a result, network D is not only a discriminator, but also an encoder. Therefore, the network D has two outputs: (1) features D_f extracted from input images; and (2) discriminative result D_d that represents the probability that input images come from the real images.

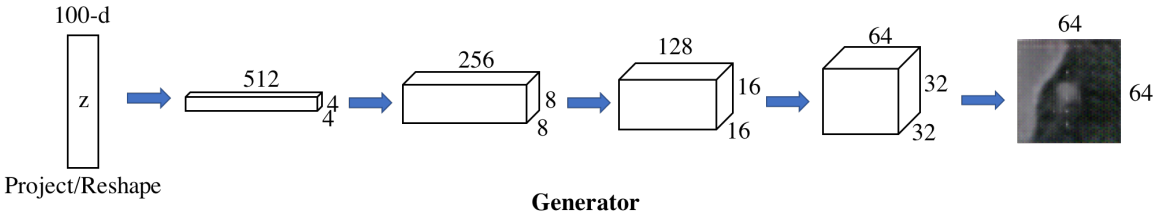

Figure 5. Network structure of the generator (G) in Forward GAN. A 100-dimensional input is converted into a 64 × 64 pixel image by four fractionally-strided convolutions. All layers in the network G use batch-normalization. ReLU activation is used in all layers except for the output layer, which uses Tanh.

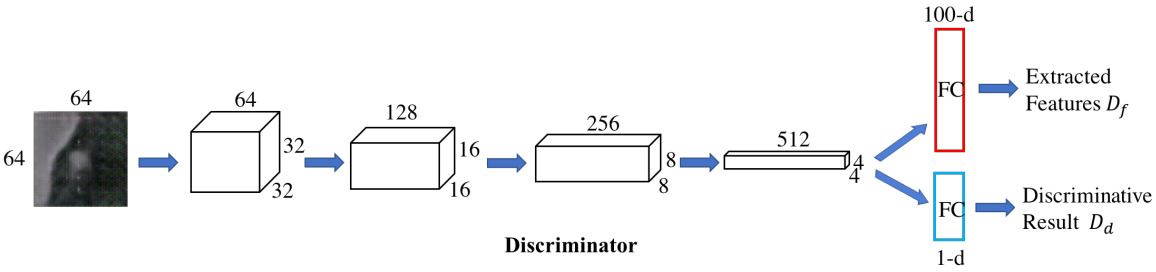

Figure 6. Network structure of the discriminator (D) in Forward GAN. A stack of strided convolutions extract features from the input, followed by two FC layers, one for discrimination and the other outputs a 100-dimensional feature. All convolutional layers in the network D use batch-normalization and LeakyReLU activation. The FC layer in cyan is followed by a sigmoid classifier.

The generator G and the discriminator D play a minimax game. The network D attempts to distinguish between real images and generated images. For each real image x, the network D outputs a discriminative result $D_d(x)$ and tries to make $D_d(x) \to 1$. Therefore, the network D should minimize $log(1 - D_d(x))$. For the generated images $G(z)$, the network D tries to make $D_d(G(z)) \to 0$. Consequently, the network D should minimize $log(D_d(G(z)))$ at the same time. The loss function for network D is:

$$L_D = log(D_d(G(z))) + log(1 - D_d(x)). \quad (1)$$

Meanwhile, the generator G tries to fool the discriminator D. For the generated images $G(D_f(x))$ and $G(z)$, the network G tries to make $D_d(G(D_f(x))) \to 1$ and $D_d(G(z)) \to 1$. Therefore, We should train G to minimize $log(1 - D_d(G(D_f(x))))$. For the the noise input z, the network G should minimize $log(1 - D_d(G(z)))$.

Besides, as shown in Figure 3, in the network G, to solve the mode collapse problem, we add another pipeline, which can obtain a mapping from the real image x to the f_x by the network D. Using f_x, we can obtain the generated image x_p by the network G. For each real image x, the network G should generate an image x_p that is indistinguishable from x. Consequently, G should satisfy the cycle consistency: $x \rightarrow D_f(x) \rightarrow G(D_f(x)) \approx x$. We can add a ℓ_2 reconstruction loss to motivate this behavior:

$$L_g = \|x - x_p\|_2^2. \tag{2}$$

Therefore, the goal of network G is to minimize the following loss function:

$$L_G = L_g + log(1 - D_d(G(D_f(x)))) + log(1 - D_d(G(z))). \tag{3}$$

The entire training pipeline of Forward GAN is described in Algorithm 1.

Algorithm 1 The training pipeline of the proposed Forward GAN algorithm.

Require: m, the batch size. θ_D, initial D network parameters. θ_G, initial G network parameters.
for number of training iteration **do**
Sample $x \sim P_r$ a batch from the real data;
$y_x, f_x \leftarrow D(x)$
Sample $z \sim P_z$ a batch of random noise;
$x_f \leftarrow G(z)$
$x_p \leftarrow G(f_x)$
$y_{x_p}, f_{x_p} \leftarrow D(x_p)$
$y_{x_f}, f_{x_f} \leftarrow D(x_f)$
$L_D \leftarrow log(y_{x_f}) + log(1 - y_x)$.
Calculate ℓ_2 reconstruction loss between x and x_p: $L_g \leftarrow \|x - x_p\|_2^2$
$L_G \leftarrow L_g + log(1 - y_{x_p}) + log(1 - y_{x_f})$
Update discriminator D by stochastic gradient descent: $\theta_D \overset{+}{\leftarrow} -\nabla_{\theta_D}(L_D)$
Update generator G by stochastic gradient descent: $\theta_G \overset{+}{\leftarrow} -\nabla_{\theta_G}(L_G)$
end for

3.1.2. Backward GAN.

Although Forward GAN can generate diverse samples, the generated images are a little blurry and noisy, and the same phenomenon occurs in DCGAN. Considering that low quality samples may have a negative impact on the classification result, we design a de-noising network Backward GAN. As shown in Figure 7, Backward GAN also includes two parts, and the structures of the network G and network D are almost the same as that in Forward GAN except that some layers are removed (e.g., FC layer). The convolutional neural network D extracts features from noisy images and eliminates unwanted noises firstly, and then the network G utilizes these features to generate clear images.

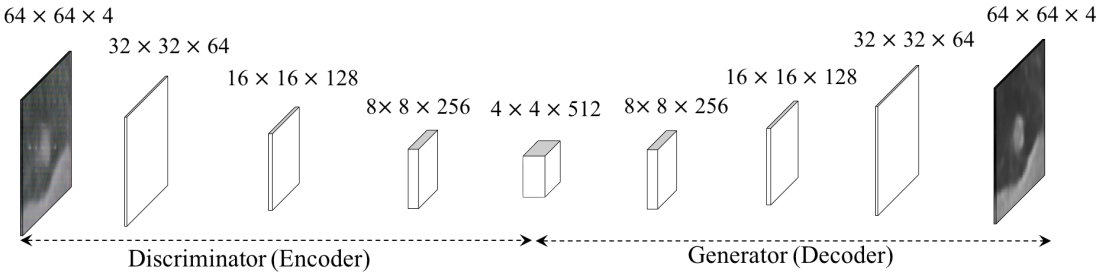

Figure 7. Network structure of the Backward GAN. Discriminator consists of a stack of strided convolutions to extract features from the noisy input. In contrast, generator includes a series of fractionally-strided convolutions. All layers use batch-normalization. LeakyReLU activation and ReLU activation are used in all layers of network D and network G, respectively.

During the training process, a pair of training data is given, including noisy images x_n and corresponding clear images x. The clear images are lung nodules images extracted from LIDC-IDRI dataset. The noisy images are obtained by adding noise to corresponding clear images manually. Backward GAN should convert x_n into the de-noised image $x_n^{'}$ so that $x_n^{'}$ is indistinguishable from x. Therefore, we train Backward GAN to minimize the Mean Absolute Deviation (MAD) between x and $x_n^{'}$. In other words, we need to minimize following loss function to optimize the network:

$$L = \frac{|x - x_n^{'}|}{N}, \tag{4}$$

where N is the total number of pixels in the image. The entire training pipeline of Backward GAN is described in Algorithm 2.

Algorithm 2 The training pipeline of the proposed Backward GAN algorithm.

Require: m, the batch size. $\theta_{D'}$, initial $D^{'}$ network parameters. $\theta_{G'}$, initial $G^{'}$ network parameters.
for number of training iteration **do**

Sample $x_n \sim P_{x_n}$ a batch from the noise data;
Sample $x \sim P_x$ a batch from the clear data;

$f_{x_n} \leftarrow D^{'}(x_n)$
$x_n^{'} \leftarrow G^{'}(f_{x_n})$
$L \leftarrow \frac{|x - x_n^{'}|}{N}$
Update discriminator $D^{'}$ and generator $G^{'}$ by stochastic gradient descent: $\theta_{D'}, \theta_{G'} \overset{+}{\leftarrow} -\nabla_{\theta_{D'}, \theta_{G'}}(L)$
end for

3.2. Multi-scale VGG16

In previous VGG16 network, all filters have the same receptive field, which leads to the loss of scale-variant information. As shown in Figure 8, to capture more features, we design a new network M-VGG16, which adds Multi-Scale Blocks (MSB) after the first four max-pooling layers of the VGG16 network.

The four MSBs in M-VGG16 have the same network structure. As shown in Figure 9, each MSB has three convolutional layers with various receptive field scale. The receptive field of first layer is 3×3, while that of other layers are 1×1. With different receptive field, MSB can capture more subtle features. In addition, the strides of each MSB are different as shown in Table 2. This allows for the extraction of different features and ensures the output size of each MSB is the same. All features are concatenated after the fifth max-pooling layer. Concatenate layer is followed by three Fully-Connected

(FC) layers: the first two FC layers have 4096 nodes, the third FC layer contains 2 nodes. The final layer is the softmax layers used to classify. The details for each MSB are shown in Table 2.

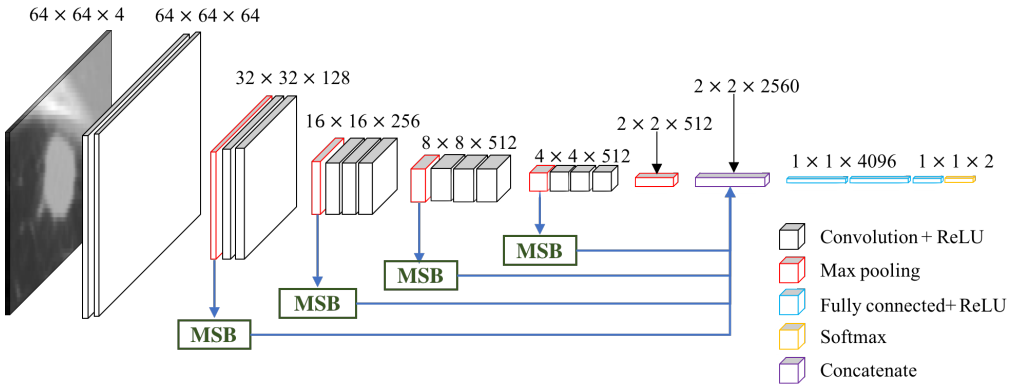

Figure 8. Network structure of the M-VGG16. In this network, except for MSBs, all filters have the same 3 × 3 receptive field, and the convolution stride is fixed to 1. All hidden layers are equipped with ReLU activation.

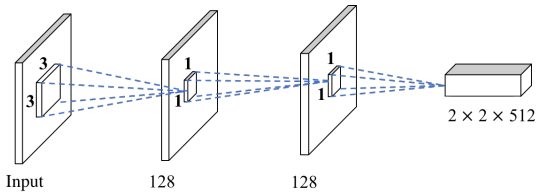

Figure 9. Network structure of the MSB. MSB is composed of three convolutional layers. The first layer has a 3 × 3 receptive field, and the last two layers have 1 × 1 receptive fields.

Table 2. Layer parameters of the multi-scale block (MSB) model.

	Layer	Input	Output	Filter Size	Strides
MSB1	conv1	32 × 32 × 64	8× 8 × 128	3	4
	conv2	8 × 8 × 128	4 × 4 × 128	1	2
	conv3	4 × 4 × 128	2 × 2 × 512	1	2
MSB2	conv1	16 × 16 × 128	4 × 4 × 128	3	4
	conv2	4 × 4 × 128	2 × 2 × 128	1	2
	conv3	2 × 2 × 128	2 × 2 × 512	1	2
MSB3	conv1	8 × 8 × 256	3 × 3 × 128	3	2
	conv2	3 × 3 × 128	2 × 2 × 128	1	2
	conv3	2 × 2 × 128	2 × 2 × 512	1	1
MSB4	conv1	4 × 4 × 512	2 × 2 × 128	3	1
	conv2	2 × 2 × 128	2 × 2 × 128	1	1
	conv3	2 × 2 × 128	2 × 2 × 512	1	1

4. Experiments and Results

4.1. Datasets

The dataset used in this paper is the LIDC-IDRI [25], which contains 1018 cases along with annotations from up to four experienced thoracic radiologists. Nevertheless, the diagnostic information is only available for nodules with a diameter larger than 3 mm. As the diagnostic information is the only way to judge the certainty of malignancy, we chose to use nodules larger than 3 mm. These nodules are rated from 1 to 5, indicating an increasing degree of malignancy (1 denotes low malignancy and 5 is high malignancy). In our paper, Rating 1 and Rating 2 are regarded as benign. Rating 4 and

Rating 5 are considered as malignant and we remove those samples with the Rating 3 for unclear malignancy. Overall, there are 353 samples including 158 benign samples and 195 malignant samples. In the training process, the original training set has 185 samples, including 83 benign samples and 102 malignant samples, and the test set has 168 samples, including 75 benign samples and 93 malignant samples. Even if all 353 samples are used for training, it is easy to cause overfitting. Therefore, data augmentation is necessary to solve this problem.

In addition, for all 353 samples, we need to extract ROI from the original lung images. In the LIDC-IDRI dataset, the coordinates of each pulmonary nodule are annotated by four professional radiologists. We read the coordinates of the nodule and take a 64×64 image with nodule coordinates as the center point. The 64×64 image is the ROI of the original CT image.

4.2. Evaluation Metrics

In this paper, we use four metrics to estimate the classification performance of our M-VGG16 model on test dataset. In our case, true positive (TP) is the number of samples correctly identified as malignant. False positive (FP) is the number of samples incorrectly identified as malignant. True negative (TN) is the number of samples correctly identified as benign. Analogously, false negative (FN) is the number of samples incorrectly identified as benign.

1. **Accuracy**: The accuracy is the ability of our model to differentiate the malignant and benign samples correctly. Mathematically, this can be defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

2. **Sensitivity**: The sensitivity represents the ability of our model to determine the malignant samples correctly. Mathematically, this can be defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

3. **Specificity**: The specificity illustrates the ability of our model to determine the benign samples correctly. Mathematically, this can be defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

4. **AUROC**: the ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system. The AUROC score is the area under the ROC curve.

4.3. Data Augmentation Performance of F&BGAN

We propose a new image synthesis method F&BGAN to enlarge dataset. However, to generate diverse images, F&BGAN also requires many training samples. To cope with this situation, we incorporate the traditional augmentation techniques for enlarging training dataset of F&BGAN. Besides, to compare F&BGAN with other data augmentation methods, we obtain a group of datasets with different size using different methods.

As shown in Figure 10, Dataset1 includes the original training data. Based on original samples, we use traditional methods, DCGAN and Forward GAN to produce 3200 new images for each class, and obtain Dataset2, Dataset3, and Dataset4, respectively. For Dataset5, we use Dataset2 to train Forward GAN and generate 6400 samples for each class. After that, we use Backward GAN to improve the image quality of Dataset5 to obtain Dataset6. Similarly, we use Dataset2 to train DCGAN and obtain Dataset7. Finally, we integrate original samples and generated samples for training. The size of each dataset is shown in Figure 10.

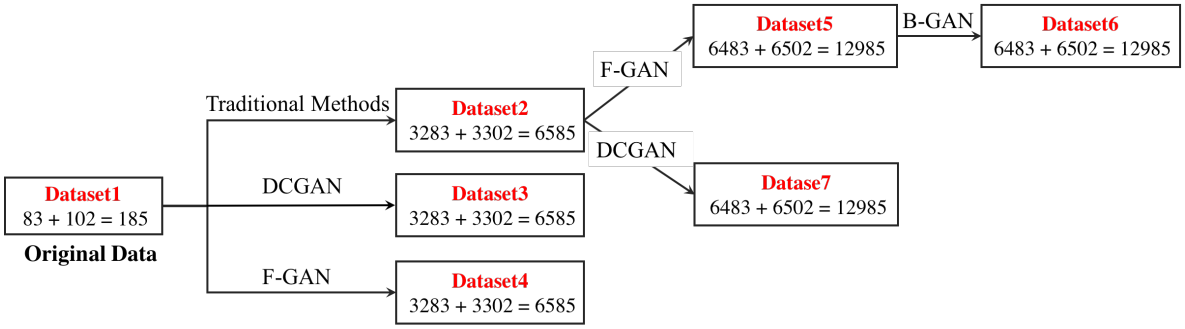

Figure 10. Datasets generated by different augmentation methods. The augmentation methods are shown on the arrow. In each box, the first number represents the number of benign samples, the second number represents the number of malignant samples and the last is the total number.

Figure 11 shows images of different dataset. The images of Dataset2 are clearer than the images of Dataset3 and Dataset4, and this is because the traditional method generates images by geometric transformation. However, the images generated by traditional method are lack of diversity. Comparing Dataset4 and Dataset5, we can draw a conclusion that more training samples lead to more realistically generated images. Besides, the images of Dataset6 are clearer than the images of Dataset5, which proves the effectiveness of BGAN. The images of Dataset5 are more realistic than the images of Dataset7, which indicates that the improvements of FGAN are effective.

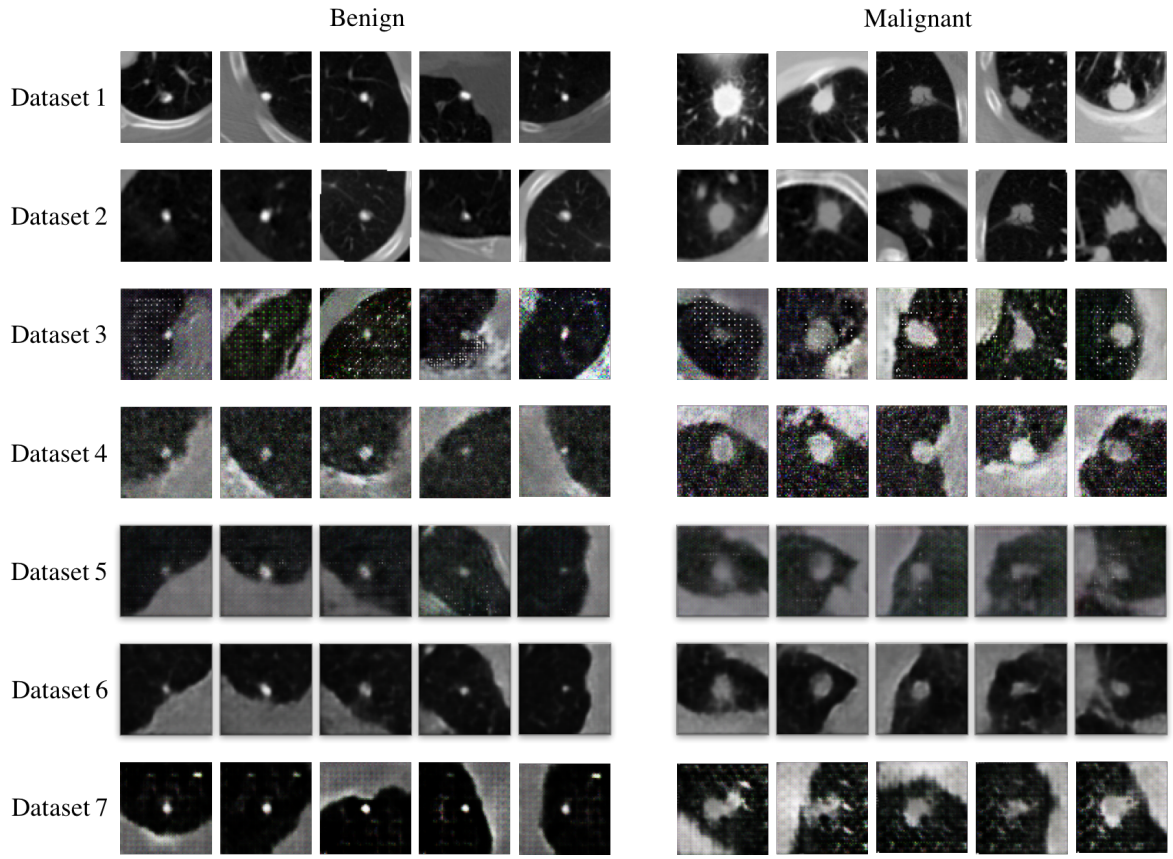

Figure 11. The generated images. The left block represents benign samples and the right block represents malignant samples.

4.4. Classification Performance of M-VGG16

To evaluate the performance of F&BGAN and M-VGG16, we design and perform the following experiments: (1) Dataset1 + VGG16; (2) Dataset1 + M-VGG16; (3) Dataset2 + M-VGG16; (4) Dataset3 + M-VGG16; (5) Dataset4 + M-VGG16; (6) Dataset5 + M-VGG16; (7) Dataset6 + M-VGG16; (8) Dataset7 + M-VGG16; (9) Dataset2 + Dataset5 + M-VGG16; and (10) Dataset2 + Dataset6 + M-VGG16.

During training, we use a 10-fold cross validation for evaluating classification performance. During training, the training dataset is randomly partitioned into 10 equal sized subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. In the testing phase, we use test set to evaluate the final model fit on the training dataset. Finally, we get experiment results with the best accuracy of 95.24%, sensitivity of 98.67%, specificity of 92.47% and AUROC of 0.980. The experiment results are shown in Tables 3 and 4. Table 4 shows directly how each component affects performance.

Table 3. Results of experiments. We evaluate the classification results of the test set with four evaluating indicators: accuracy, sensitivity, specificity and AUROC. The number in the second column represents the size of training set in the each experiment. The third column is the classification method.

Num	Training Set	Classification	ACC (%)	SEN (%)	SPE (%)	AUROC
1	Dataset1(185)	VGG16	72.62	58.67	83.87	0.835
2	Dataset1(185)	M-VGG16	88.09	85.33	90.32	0.954
3	Dataset2(6585)	M-VGG16	92.86	96.00	90.32	0.982
4	Dataset3(6585)	M-VGG16	90.48	88.00	92.47	0.966
5	Dataset4(6585)	M-VGG16	91.67	89.33	93.55	0.968
6	Dataset5(12985)	M-VGG16	92.86	92.00	93.55	0.967
7	Dataset6(12985)	M-VGG16	93.45	91.04	95.29	0.980
8	Dataset7(12985)	M-VGG16	92.26	90.67	93.55	0.973
9	Dataset2 + Dataset5(19385)	M-VGG16	94.05	98.67	90.32	**0.984**
10	Dataset2 + Dataset6(19385)	M-VGG16	**95.24**	**98.67**	**92.47**	0.980

Table 4. Effects of various design choices and components on performance. The check mark represents that the component is used in experiment. The red check mark means FGAN, BGAN and DCGAN are trained using Dataset2 not the original Dataset1. Tra is the abbreviation of traditional augmentation method.

Num	Tra	DCGAN	FGAN	BGAN	M-VGG16	ACC (%)	SEN (%)	SPE (%)	AUROC
1						72.62	58.67	83.87	0.835
2					√	88.09	85.33	90.32	0.954
3	√				√	92.86	96.00	90.32	0.982
4		√			√	90.48	88.00	92.47	0.966
5			√		√	91.67	89.33	93.55	0.968
6			√		√	92.86	92.00	93.55	0.967
7			√	√	√	93.45	91.04	95.29	0.980
8		√			√	92.26	90.67	93.55	0.973
9	√		√		√	94.05	98.67	90.32	**0.984**
10	√		√	√	√	**95.24**	**98.67**	**92.47**	0.980

With the same training dataset, the result of Experiment 2 is much better than Experiment 1, which proves that the M-VGG16 classifier outperforms traditional VGG16. The results of Experiments 3–5 show that our method Forward GAN is superior to DCGAN. However, the classification performance of GAN-based methods is not as good as traditional data augmentation methods, which is probably because that images generated by traditional methods are more realistic. The result of Experiment 6 is better than Experiment 5, and this implicitly proves that the more training samples the more realistic the generated images. The training samples in Experiment 6 are generated by Forward GAN, while training samples in Experiment 7 are generated by F&BGAN. The experiment results indicate that

proposed Backward GAN is necessary and effective in removing noise and improving image quality. Analogously, the results of Experiments 9 and 10 can also prove the effectiveness of BGAN. The results of Experiments 6 and 8 demonstrate once again that FGAN has better performance than DCGAN in image generation. Compared with the Experiment 1, the accuracy of Experiment 10 is increased by 22.78%, which powerfully proves the effectiveness of F&BGAN and M-VGG16.

4.5. Performance of MSB

In addition, we have designed a set of experiments to verify the effectiveness of MSBs. The first experiment uses VGG16 without any MSB. The second experiment uses only the first one MSB. The third experiment uses the first two MSBs. The fourth experiment uses the first three MSBs. The last experiment uses four MSBs. All experiments use the same training dataset (Dataset2 + Dataset6) and the same testing dataset. In addition, 10-fold cross validation is used during the training process.

The experimental results are shown in Table 5. The results illustrate that, the more MSB is used, the better is the classification performance. In addition, compared with VGG16, the accuracy of M-VGG16 is improved by 4%, which proves that MSBs are effective to enhance the classification performance.

Table 5. Comparison of classification performance using different numbers of MSBs.

Experiment	Number	ACC (%)	SEN (%)	SPE (%)	AUROC
1	0	91.07	96.00	87.10	0.976
2	1	92.26	96.00	89.25	0.977
3	2	92.86	97.34	89.25	0.978
4	3	94.05	98.67	90.32	0.978
5	4	**95.24**	**98.67**	**92.47**	**0.980**

4.6. Comparison with Other Classification Methods

To further evaluate the classification performance of M-VGG16, we compare M-VGG16 with other traditional methods and neural network methods. All methods use the same training dataset (Dataset2 + Dataset6) to train the classification model. Besides, the same test dataset is used to evaluate those models. In addition, 10-fold cross validation is used during the training process. We measure the performance of different classification methods with the accuracy, specificity, sensitivity and AUROC. The experimental results are shown in Table 6.

Table 6. Comparison between our method and the state-of-the-art in classification.

Approach	Methods	ACC (%)	SEN (%)	SPE (%)	AUROC
Traditional	KNN [29]	68.45	66.67	69.89	0.683
	Softmax	88.10	89.34	87.10	0.882
	SVM [28]	89.88	88.00	91.40	0.897
Neural Network	CNN [30]	87.50	93.33	82.80	0.935
	GoogLeNet [31]	89.88	97.33	83.87	0.950
	ResNet [32]	93.45	98.67	89.25	0.979
	VGG16 [23]	91.07	96.00	87.10	0.976
	Our Method(M-VGG16)	**95.24**	**98.67**	**92.47**	**0.980**

The experimental results show that, although the same dataset is used, the performance of different classification methods varies greatly. The neural network methods are superior to the traditional methods, except for CNN. This may be due to the CNN we use has a relatively simple network structure. Only three convolutional layers and two full-connected layers are used in CNN. For neural network methods, our M-VGG16 network outperforms others. Furthermore, M-VGG16

is superior to ResNet while VGG16 is inferior to ResNet. It is remarkable that M-VGG16 is 4% more accurate than the traditional VGG16 network with the help of MSB.

4.7. Comparison with the State-Of-the-Art

Furthermore, we compare the results of our method with the results of other related methods, including traditional methods and neural network-based methods. As shown in Table 7, we statistically measure the performance of different methods with the accuracy, specificity, sensitivity and AUROC.

Table 7. Classification performance comparison with the state-of-the-art with reference to the classification of lung nodules. Tra is the abbreviation of traditional methods. NN is the abbreviation of neural network based methods. The test column indicates the size of the test set for each method.

	Research	**Database**	**Test**	**ACC** (%)	**SEN** (%)	**SPE** (%)	**AUROC**
Tra	Farag et al. [3]	ELCAP (294)	-	-	86.0	86.0	-
	Orozco et al. [4]	NBIA-ELCAP (113)	75	84.00	83.33	83.33	-
	Krewer et al. [6]	LIDC-IDRI (33)	-	87.88	85.71	89.47	-
	Parveen and Kavitha [5]	Private (3278)	1639	-	91.38	89.56	-
NN	Hua et al. [8]	LIDC (2545)	-	-	73.40	82.80	-
	Kumar et al. [10]	LIDC (4323)	432	75.01	83.35	-	-
	Shen et al. [12]	LIDC-IDRI (16764)	275	86.84	-	-	-
	Cheng et al. [11]	LIDC (10133)	140	95.6	92.4	98.9	0.989
	Shen et al. [7]	LIDC-IDRI (1375)	275	87.14	77.0	93.0	0.93
	Kwajiri and Tezuka [13]	LIDC-IDRI (748)	299	-	89.50	89.38	-
	Abbas [14]	LIDC-IDRI (3250)	2112	-	88	80	0.89
	Da Silva et al. [15]	LIDC-IDRI (21631)	1343	94.78	94.66	95.14	0.949
	Our method	LIDC-IDRI (19553)	168	**95.24**	**98.67**	**92.47**	**0.980**

Among all the methods, Cheng et al. [11] has the best accuracy of 95.6%, and our method has the second-best accuracy of 95.24%, with the difference of 0.36%; however, the sensitivity of our method is better, with the difference of 6.27%. It is worth noting that the sensitivity shows the performance to classify correctly malignant nodules, which is more important in CAD system. Besides, Cheng et al. [11] selected 700 malignant and 700 benign samples from LIDC datasets, while we only chose 158 benign samples and 195 malignant samples due to different selection criteria and methods. The size of our original samples is relatively small. Da Silva et al. [15] used more samples than us, i.e., 21,631 samples. Nevertheless, their accuracy is lower than ours.

5. Conclusions

In this paper, a new data augmentation method F&BGAN is proposed to tackle the limited data problem. In addition, we design a multi-scale VGG16 to classify lung nodules as benign or malignant. By using F&BGAN and M-VGG16, the accuracy of lung nodules classification is increased by 22.78%, which demonstrates the feasibility of F&BGAN in generating medical images and the effectiveness of M-VGG16 in classifying malignant and benign nodules. In the future, we will further boost the classification performance by combining different CNNs with M-VGG16 by referring to the work of Nanni et al. [33]. In addition, we plan to extend our work to other diseases.

Author Contributions: D.Z. (Defang Zhao) conceived and designed the experiments; D.Z. (Defang Zhao) performed the experiments; D.Z. (Defang Zhao), D.Z. (Dandan Zhu) and G.Z. analyzed the data; Y.L., J.L. and D.Z. (Dandan Zhu) contributed reagents/materials/analysis tools; D.Z. (Defang Zhao) wrote the paper.

Funding: This work was supported by the General Program of National Natural Science Foundation of China (NSFC) under Grant No. 61572362. This research was also partially supported by the General Program of National Natural Science Foundation of China (NSFC) under Grant No. 81571347, Fundamental Research Funds for the Central Universities under Grant No. 22120180012, National Natural Science Foundation of China (NSFC) under Grant No. 61806147 and Shanghai Natural Science Foundation under Grant No. 18ZR1441200.

Acknowledgments: The authors are grateful for the comments and reviews from the reviewers and editors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jemal, A.; Siegel, R.; Ward, E.; Murray, T.; Samuels, A.; Tiwari, R.C.; Ghafoor, A.; Feuer, E.J.; Thun, M.J. Cancer statistics. *CA Cancer J. Clin.* **2008**, *58*, 71–96. [CrossRef] [PubMed]
2. Ramaswamy, S.; Truong, K. Pulmonary Nodule Classification with Convolutional Neural Networks. *Comput. Math. Methods Med.* **2016**. [CrossRef]
3. Farag, A.; Ali, A.; Graham, J.; Farag, A.; Elshazly, S.; Falk, R. Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose CT scans of the chest. In Proceedings of the 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Chicago, IL, USA, 30 March–2 April 2011; pp. 169–172.
4. Orozco, H.M.; Villegas, O.O.V.; Domínguez, H.J.O.; Domínguez, H.D.J.O.; Sanchez, V.G.C. Lung nodule classification in CT thorax images using support vector machines. In Proceedings of the 2013 12th Mexican International Conference on Artificial Intelligence (MICAI), Mexico City, Mexico, 24–30 November 2013; pp. 277–283.
5. Parveen, S.S.; Kavitha, C. Classification of lung cancer nodules using SVM Kernels. *Int. J. Comput. Appl.* **2014**, *95*, 975–8887.
6. Krewer, H.; Geiger, B.; Hall, L.O.; Goldgof, D.B.; Gu, Y.; Tockman, M.; Gillies, R.J. Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography. In Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Manchester, UK, 13–16 October 2013; pp. 3887–3891.
7. Shen, W.; Zhou, M.; Yang, F.; Yu, D.; Dong, D.; Yang, C.; Zang, Y.; Tian, J. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recogn.* **2017**, *61*, 663–673. [CrossRef]
8. Hua, K.L.; Hsu, C.H.; Hidayati, S.C.; Hidayati, S.C.; Cheng, W.H.; Chen, Y.J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets Ther.* **2015**, *8*, 2015–2022.
9. Dandıl, E.; Çakiroğlu, M.; Ekşi, Z.; Özkan, M.; Kurt, Ö.K.; Canan, A. Artificial neural network-based classification system for lung nodules on computed tomography scans. In Proceedings of the 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPar), Tunis, Tunisia, 11–14 August 2014; pp. 382–386.
10. Kumar, D.; Wong, A.; Clausi, D.A. Lung nodule classification using deep features in CT images. In Proceedings of the 2015 12th Conference on Computer and Robot Vision (CRV), Halifax, NS, Canada, 3–5 June 2015; pp. 133–138.
11. Cheng, J.Z.; Ni, D.; Chou, Y.H.; Qin, J.; Tiu, C.M.; Chang, Y.C.; Huang, C.S.; Shen, D.; Chen, C.M. Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.* **2016**, *6*, 24454. [CrossRef] [PubMed]
12. Shen, W.; Zhou, M.; Yang, F.; Yang, C.; Tian, J. Multi-scale convolutional neural networks for lung nodule classification. In Proceedings of the International Conference on Information Processing in Medical Imaging, Cham, Switzerland, 28 June–3 July 2015; pp. 588–599.
13. Kwajiri, T.L.; Tezukam T. Classification of Lung Nodules Using Deep Learning. *Trans. Jpn. Soc. Med. Biol. Eng.* **2017**, *55*, 516–517.
14. Abbas, Q. Lung-Deep: A Computerized Tool for Detection of Lung Nodule Patterns using Deep Learning Algorithms. *Lung* **2017**, *8*. [CrossRef]
15. Da Silva, G.L.F.; da Silva Neto, O.P.; Silva, A.C.; Gattass, M. Lung nodules diagnosis based on evolutionary convolutional neural network. *Multimed. Tools Appl.* **2017**, *76*, 19039–19055. [CrossRef]
16. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Mach. Learn.* **2014**, 2672–2680, arXiv:1406.2661.
17. Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, C.; Greenspan, H. Synthetic data augmentation using GAN for improved liver lesion classification. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 289–293.

18. Guibas, J.T.; Virdi, T.S.; Li, P.S. Synthetic Medical Images from Dual Generative Adversarial Networks. *arXiv* **2017**, arXiv:1709.01872.
19. Chuquicusma, M.J.M.; Hussein, S.; Burt, J.; Bagci, U. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 240–244.
20. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2016**, arXiv:1511.06434.
21. Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; Courville, A. Adversarially learned inference. *arXiv* **2016**, arXiv:1606.00704.
22. Donahue, J.; Krähenbühl, P.; Darrell, T. Adversarial feature learning. *arXiv* **2016**, arXiv:1605.09782
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
24. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–27 September 1999; pp. 1150–1157.
25. Armato, S.G., III; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [PubMed]
26. Farag, A.; Elhabian, S.; Graham, J.; Farag, A.; Falk, R. Toward precise pulmonary nodule descriptors for nodule type classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2010; pp. 626–633.
27. Ojala, T.; Pietikäinen, M.; Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.* **1996**, *29*, 51–59. [CrossRef]
28. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
29. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in neural information processing systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
31. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Nanni, L.; Ghidoni, S.; Brahnam, S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognit.* **2017**, *71*, 158–172. [CrossRef]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).