BIOCYBERNETICS AND BIOMEDICAL ENGINEERING 41 (2021) 1243–1257

Available at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning

Ioannis D. Apostolopoulos [a,*], *Nikolaos D. Papathanasiou* [b], *George S. Panayiotakis* [a]

[a] *Department of Medical Physics, School of Medicine, University of Patras, 26504 Patras, Greece*
[b] *Department of Nuclear Medicine, University Hospital of Patras, 26504 Patras, Greece*

ARTICLE INFO

Article history:
Received 17 May 2021
Received in revised form
23 August 2021
Accepted 23 August 2021
Available online 2 September 2021

Keywords:
Lung nodule classification
Data augmentation
Generative adversarial networks
Medical image classification
Semi-supervised learning

ABSTRACT

The pulmonary nodules' malignancy rating is commonly confined in patient follow-up; examining the nodule's activity is estimated with the Positron Emission Tomography (PET) system or biopsy. However, these strategies are usually after the initial detection of the malignant nodules acquired from the Computed Tomography (CT) scan. In this study, a Deep Learning methodology to address the challenge of the automatic characterisation of Solitary Pulmonary Nodules (SPN) detected in CT scans is proposed.

The research methodology is based on Convolutional Neural Networks, which have proven to be excellent automatic feature extractors for medical images. The publicly available CT dataset, called Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI), and a small CT scan dataset derived from a PET/CT system, is considered the classification target. New, realistic nodule representations are generated employing Deep Convolutional Generative Adversarial Networks to circumvent the shortage of large-scale data to train robust CNNs. Besides, a hierarchical CNN called Feature Fusion VGG19 (FF-VGG19) was developed to enhance feature extraction of the CNN proposed by the Visual Geometry Group (VGG). Moreover, the generated nodule images are separated into two classes by utilising a semi-supervised approach, called self-training, to tackle weak labelling due to DC-GAN inefficiencies.

The DC-GAN can generate realistic SPNs, as the experts could only distinguish 23% of the synthetic nodule images. As a result, the classification accuracy of FF-VGG19 on the LIDC-IDRI dataset increases by +7%, reaching 92.07%, while the classification accuracy on the CT dataset is increased by 5%, reaching 84,3%.

* Corresponding author.
E-mail address: ece7216@upnet.gr (I.D. Apostolopoulos).

https://doi.org/10.1016/j.bbe.2021.08.006

1. Introduction

Computer Tomography (CT) constitutes the traditional procedure for lung nodule detection. In contrast, reliably distinguishing between benign and malignant nodules involves either patient follow up, biopsy, or FDG uptake, computed by PET/CT systems [1]. In addition, the nodule suspiciousness rating is accurately defined with Positron Emission Tomography (PET) imaging [2].

The traditional pipeline for the computer-aided classification of lung nodules based on medical imaging involves handcrafted feature extraction and classification with the aid of Machine Learning algorithms [3]. However, due to the shortage of large–scale image data, researchers investigated methods and techniques to improve the classification accuracy without additional data. Recently, the public availability of large CT datasets directed the research towards Deep Learning. Deep Learning with Convolutional Neural Networks (CNNs), alludes to various Machine Learning methods utilising larger databases [4], which is based on automatic deep feature extraction.

The performance of the CNNs depends on the data size, and therefore, several data augmentation techniques have been proposed. Conventional data augmentation methods are mainly geometric transformations, such as flip, distort, crop, and rotate. Recently, Goodfellow proposed Generative Adversarial Networks (GAN) [5] to generate synthetic images. However, the generated images have no direct connection with the original image due to their creation by an independent network. The original and generated images similarity is not easily distinguishable by the human eye.

Besides, data scarcity can be circumvented by avoiding the training of CNNs from scratch and employing pre-trained CNN models. Those CNNs are initially trained on large sets of images, while they have learned to extract high-level features. Transferring this knowledge of the CNNs is a procedure called transfer learning. One scheme of transfer learning is fine-tuning, wherein a proportion of the borrowed network's architecture can learn new features from the target images, while the rest comes with fixed weights defined from the initial training process.

In this paper, the problem of the automatic classification of SPNs is considered. The targets for accurate classification are the publicly available LIDC-IDRI [6] dataset and a smaller CT scan dataset obtained from a PET/CT scanner.

The intention of this study is three-fold; firstly, to evaluate the performance of Deep Convolutional Generative Adversarial Networks (DC-GAN), which are employed to generate new lung nodule representations, aiming to expand the training datasets with new realistic data. Secondly, to investigate the effectiveness of the semi-supervised method in labelling the generated nodule images, since the images were weakly labelled by the DC-GAN and could not be of use to enhance the training sets. Thirdly, to propose an innovative fine-tuned modification of the CNN developed by the Visual Geometry Group and named after it (VGG), one of the most famous models for classification tasks involving medical images. Its structure was modified to achieve multi-level feature extraction. For the first task, a Deep Convolutional Generative Adversarial Network (DC-GAN) generates new images based on the two datasets, as mentioned. The evaluation of the efficiency of the DCGAN is performed in two stages. Two Nuclear Medicine doctors were assigned to distinguish the fake from the real nodules for the first stage. For the second stage, the generated nodule images are utilised by the classification CNNs of the experiment to expand the training sets and obtain better classification accuracy on either dataset. As far as the second intention of the study is concerned, the DC-GAN was trained to generate new images for each of the classes (benign/malignant). The new images are also labelled using a Semi-Supervised technique called self-training to evaluate the effectiveness of the DC-GAN for the generation of realistically labelled images. Finally, the two labelling strategies are compared based on the classification accuracy of the CNNs. The state-of-art pre-trained CNN, named VGG19 [7], is employed for the classification task. VGG has been a successive CNN for similar tasks, as suggested by recent studies [8,9]. A new VGG19 modification is applied in this study to extract more significant features during the late convolutional processes. The proposed CNN is called Feature Fusion – VGG19 (FF-VGG19).

2. Related work

GANs were recently employed to generate fake lung nodule images to improve the accuracy and reduce the false positives in lung nodule detection and segmentation tasks. Some noteworthy studies are presented.

Chuquicusma et al. [10] used unsupervised learning with Deep Convolutional-Generative Adversarial Networks (DC-GANs) to generate lung nodule samples realistically. For the evaluation of the proposed model, two radiologists were asked to distinguish real from fake images. The results showed that the generated samples managed to deceive radiologists, which underlined the possibility of adopting the same strategy for detection, segmentation, and classification tasks.

Moreover, Javaid and Lee [11] proposed a pure GAN to generate whole CT images for lung representation. The generated CT images have excellent global and local features of an actual CT image and can augment the training datasets for effective learning.

Jin et al. [12] developed a 3D GAN that effectively learns lung nodule property distributions in 3D space, referred to as CGAN. The generator's input is a volume to embed the nodules within their background context, while the nodule's central part has been erased. They also proposed a multi-mask reconstruction loss. The results demonstrate that the Progressive Holistically Nested Neural Network (PHNN), presented by Harisson et al. in [13], improved the effectiveness in the segmentation task, especially for nodules adjoining the lung boundary.

Han et al. [14] proposed a 3D Multi-Conditional GAN (MCGAN) to generate realistic nodules placed naturally on

CT images to boost the sensitivity of detection of 3D objects. Their method adopts two discriminators for conditioning: the context discriminator, which learns to classify the real and the synthetic nodule/surrounding pairs with noise box-centred surroundings; The results demonstrated that the 3D Convolutional Neural Network-based detection could achieve higher sensitivity under any nodule size/attenuation, at fixed False Positive rates, and overcome the medical data paucity with the MCGAN-generated realistic nodules.

Tang et al. [15] proposed a framework of stacked GANs, called SGAN, to generate lung nodules to improve the segmentation task. The first GAN's task is noise reduction, whereas the second GAN generates a higher resolution image with enhanced boundaries and high contrast. Two conventional segmentation methods called GrabCut [16] and modern holistically nested network (HNN) [17] were utilised for the inspection of SGAN's contribution. Experimental results on the Deep Lesion [18] dataset demonstrated that the SGAN enhancements could push the performance of GrabCut over HNN, trained on original images.

Bi et al. [19] proposed multi-channel generative adversarial networks (M-GAN) for PET image synthetization. The M-GAN approach can capture feature representations with high semantic information based on the adversarial learning concept. The M-GAN can take the input from the annotation (label) to synthesise regions of high uptake, e.g., tumours and from the computed tomography (CT) images, to constrain the appearance consistency, based on the CT derived anatomical information in a single framework, and to output the synthetic PET images directly. In their work, the authors used 50 lung cancer images for the generation of new instances. Using this method, the authors tested the performance of state-of-the-art segmentation network FCN [20] by utilising either actual data or generated data. The FCN trained with synthetic PET images performed competitively to the same model trained with authentic PET images.

The GAN strategy was also adopted for lung nodule classification. Although the generated nodules are not characterised as to their malignancy by any actual means, their incorporation within the training sets enhanced the classification accuracy of the CNNs.

Zhao et al. [9] proposed Forward and Backward GAN (F&BGAN) to generate high-quality synthetic medical images. The Forward GAN (FGAN) generated diverse images, while the Backward GAN (BGAN) improved the quality of the generated images. A hierarchical learning framework, called a multi-scale VGG16 (M-VGG16) network, was proposed to extract discriminative features from alternating stacked layers. The methodology was evaluated on a small set of the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset, with the best accuracy of 95.24%, the sensitivity of 98.67%, the specificity of 92.47% and the Area Under ROC curve (AUROC) of 0.980. The experimental results demonstrated the feasibility of F&BGAN in generating medical images and the effectiveness of M-VGG16 in classifying malignant and benign nodules.

Onishi et al. [21] used GANs to generate additional images from CT images of 60 cases with confirmed pathological diagnosis by biopsy. They developed a CNN named DCNN to investigate the classification performance between benign and malignant nodules, with actual and synthetic nodules. The DCNN is trained to utilise images generated by the GAN and is fine-tuned, making full use of the actual nodule images to allow the DCNN to distinguish between benign and malignant nodules. This pre-training and the fine-tuning process could differentiate between 66.7% of benign nodules and 93.9% of malignant nodules. These results indicate that the proposed method improves the classification accuracy by approximately 20% compared to training utilising only the original images.

Yang et al. [22] proposed a class-aware adversarial synthesis framework to produce new nodules embedded in CT images. The framework is built with one generator and two class-aware discriminators. The trained networks can generate diverse nodules by conditioning the random latent variables, and the target nodule labels gave the same context. The evaluation was done using the LIDC – IDRI dataset. The results demonstrated that combining the actual image patches and the synthetic lung nodules in the training set can improve the mean AUC classification score across different network architectures by 2%.

GANs are not thoroughly explored to aid in the classification task since the generated nodule images lack accurate labelling to expand the training datasets. Thus, one of the unique characteristics of this work lies in proposing a semi-supervised approach to simultaneously label the generated nodules and utilise them to estimate the malignancy rating of images depicting real nodules.

3. Materials and methods

In this section, the datasets utilised for this study, the structure of the GAN developed to generate synthetic SPN images, and the proposed CNN for the classification tasks are described.

3.1. *Datasets*

3.1.1. *CT*

The present study intends to develop a methodology capable of working with PET/CT scanner CT images. PET/CT scanner is essential technology of the authors' department. The CT dataset was recorded at the Clinical Sector of the Department of Nuclear Medicine of the University Hospital of Patras. 112 CT scans from the PET/CT scanner were examined, consisting of 172 SPNs. Malignant nodules are 85, and benign nodules are 87. Labelling had been done by the physicians or radiologists using either (a) biopsy results, (b) FGD consumption, computed by the PET-CT Imaging technology (General Electric Healthcare: Discovery iQ3 sl16), or (c) patient follow-up. The nodules were extracted in a 2D format to fit a size of 32×32 pixels. The slice in which the nodule appears in its largest size was selected. An example of benign and malignant cases is given in Fig. 1. The dataset was incorporated into the training of the DC-GAN network to offer greater variety in pulmonary nodule features. The reader should note that only CT images are extracted from the PET/CT scanner data. Henceforth, the CT images of this scanner are defined as CT dataset.

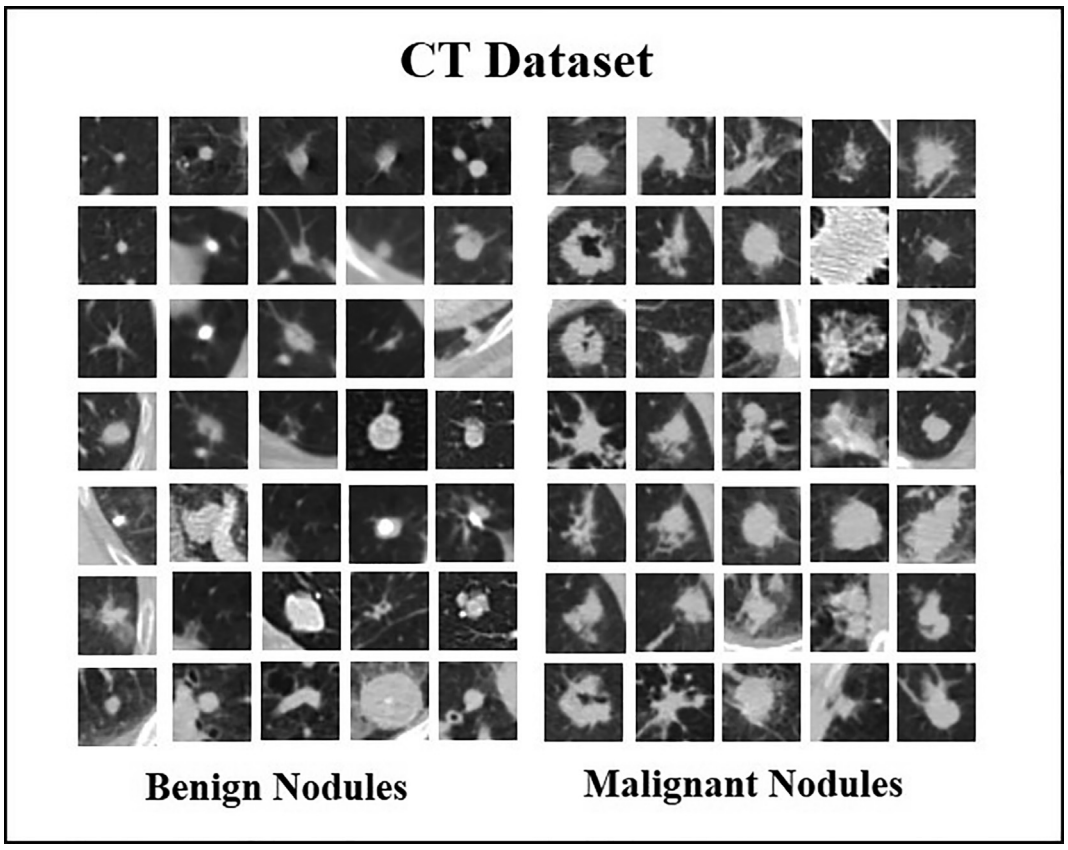

Fig. 1 – Solitary Pulmonary Nodule images from the CT dataset.

3.1.2. *LIDC-IDRI dataset*

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset consists of 1018 diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions by four radiologists. This dataset was initiated by the National Cancer Institute (NCI) [23]. For each volume, the malignancy rating is rated by four radiologists, with possible values from 1 to 5. Due to weak labelling, nodules with a mean malignancy rating between 2.5 and 3.5 were excluded from the dataset. Besides, nodules larger than 30 mm, or smaller than 3 mm were not incorporated. The final dataset consists of 620 benign nodules and 616 malignant. The nodules were extracted in a 2D format to fit a size of 32x32 pixels. The slice in which the nodule appears in its largest size was selected.

3.2. *Deep Convolutional Generative Adversarial Network (DC-GAN)*

Generative Adversarial Networks (GANs) are a unique deep Neural Network architecture. In this section, an attempt is made to describe the operation of the GAN networks without a deep analysis of their mathematical process. At the same time, the parameters selected to carry out the specific experiment are described. The reader should note that the conventional DC-GAN scheme was selected and tuned to produce realistic nodule representations for this experiment.

The selected DC-GAN framework consists of two independent CNNs, as shown in Fig. 2. The first network acts as an image generator and is named Generator (G). The second network distinguishes between authentic and generated images and is referred to as Discriminator (D). The input of the generator is an arbitrary 500-dimensional space, denoted as z. The output of G is a synthetic image, symbolised as X_g. The discriminator's input, besides the fake image X_g, is also an actual image, X_r. The output of D is an assigned probability of the image being fake or real.

Generator: The input to G, symbolised as z, is pure random noise sampled from a prior distribution p(z), commonly chosen as a Gaussian or a uniform distribution for simplicity. The recommended architecture of the generator is illustrated in Fig. 3. Four sets of 2D transposed convolutional layers [24] are applied to a random 500-dimensional input to produce the 32×32 image gradually. The input noise is transformed into an image with the transposed convolutions, so the transposed convolutions are also called deconvolutions. Between each convolutional set, the weights of the features are normalised by applying Batch Normalisation [25] layers, while 50% of the features are discarded, by applying Dropout layers [26], to prevent the generator from learning too specific information. Both methods have proven to be vital to produce realistic images [27]. The layers are activated across the entire network by the Leaky ReLU [28] function, which ensures non-linearity and negative weight distribution. Finally, the last layer is activated by the tanh function to ensure that the pixel values are normalised to the desired interval [0, 1].

Discriminator: The Discriminator, as shown in Fig. 4, is a deep CNN trained to identify real and synthetic images. The

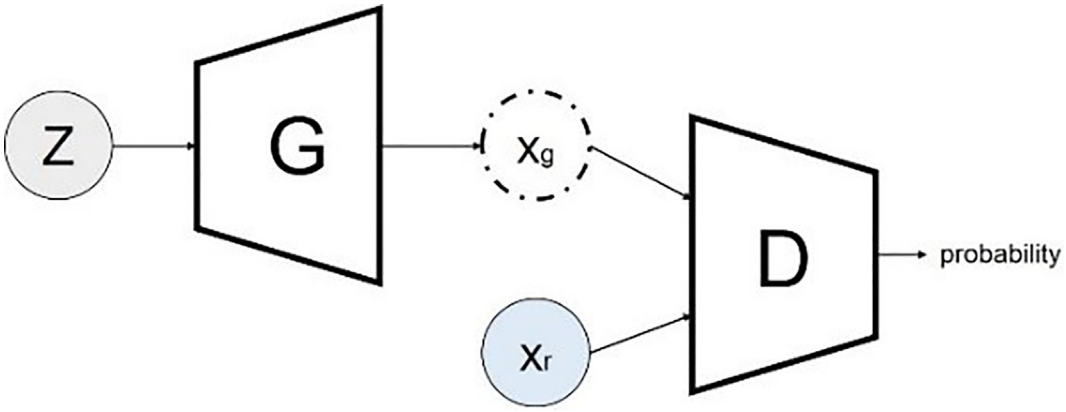

Fig. 2 – The scheme of the DC-GAN.

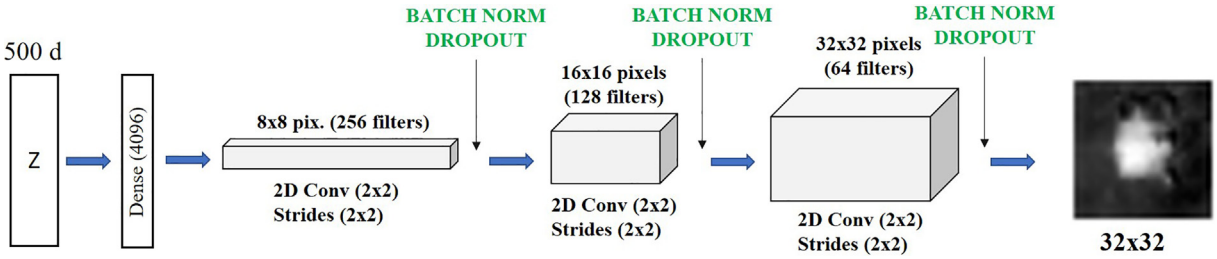

Fig. 3 – The architecture of the Generator. A 500-dimension latent space is converted into a 32 × 32 image by applying three transposed convolutions.

input image is being processed by five sets of convolutional layers of 3x3 filters. The initial image is downsampled gradually to 4 × 4 pixel representation, each incorporating features extracted by the filters applied. Every layer is activated by the Leaky ReLU function, while the robustness of the discrimination task is improved with the dropout method. Finally, the extracted features are flattened to a one-dimensional vector, processed by densely connected layers and a Softmax classifier to classify the image.

3.2.1. *DC-GAN functionality*

At each iteration, the two CNNs are co-operating to produce realistic images. The generator has sets of generated images, while the discriminator attempts to distinguish them from a set of given real images. The better the discriminator recognises the false images, whereas the generator learns to produce images not perceived by the discriminator, the better the result. The reader should note that the generator CNN of the DC-GAN is operating without using any image data. Images are given to the discriminator only to help it distinguish between fake and real inputs. GAN-generated images are neither derivatives from nor similar to the real images.

This process can last for a large number of epochs but can also fail. Common reasons leading to failure are a) the model parameters oscillate and never converge to a stable condition, which may happen due to high learning rates, or the big difference in capabilities between the two networks, b) the generator discovers a flaw in the discriminator and tends to produce limited varieties of samples, which are not recognised by the discriminator, which is known as mode collapse, and c) the discriminator gets too successful at distinguishing the fake images that the generator gradient vanishes, which is known as the diminished gradient, therefore to prevent these undesirable situations, several methods have been pro-

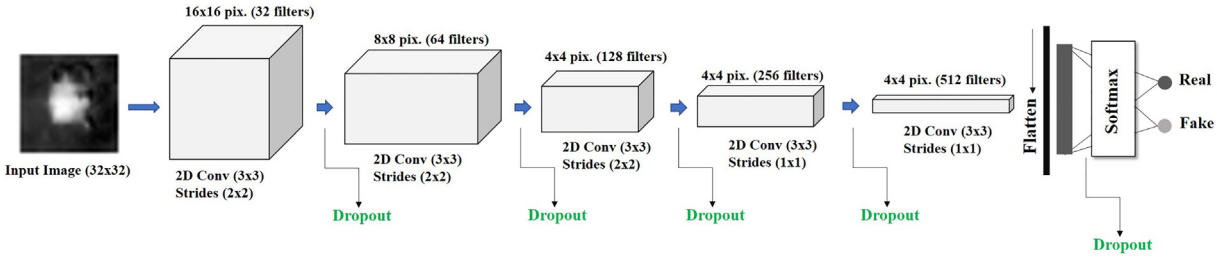

Fig. 4 – The architecture of the Discriminator. Through five 2D Convolutional layers of size 3 × 3, the input image is transformed to a 4 × 4 × 512 vector before it is flattened for the classification by the Softmax classifier.

posed. In the particular application of DC-GAN, the architectures of the networks were selected so that the one CNN is not superior to the other, and their learning rates were adjusted to a fixed value that does not cause instability.

In this experiment, the DC-GAN was compiled utilising an optimisation method, commonly known as Adam [29], which computes individual learning rates for each network parameter while achieving rapid training. Also, there is the possibility to tune the parameters called beta1 and beta2, which affect the exponential decay rate for the first and second-moment estimates accordingly. After several experiments, the generator, discriminator, and DC-GAN parameters that produce the desired images were defined and presented in Table 1.

After the training process, which lasts 500 epochs, the DC-GAN was asked to produce 15.000 lung nodule image patches equally distributed between the two classes. The generated dataset is referred to as GAN.

3.3. *Feature Fusion VGG19 (FF-VGG19) for classification*

Based on the architecture of VGG19, a modification is proposed to obtain more information from the deep feature-extracting layers. As illustrated in Fig. 5, FF-VGG19 consists of 19 Convolution Layers and 4 Max Pooling Layers uniformly distributed across the network (as the initial structure of VGG19). The outputs of the last three Max Pooling layers are connected to a series of the following layers: (a) Batch Normalisation, (b) Dropout, and (c) Global Average Pooling. Each of these series is referred to as BD. In this way, information extracted during the sequential image process is retained and not further processed. Furthermore, to reduce the number of the trainable parameters of FF-VGG19 and retain the layers' early weights, to allow the extraction of early features (edges, shapes), all the bottom 12 layers of VGG19 are untrainable.

The idea behind the extra BD layers is based on the results of Zhao et al. [9]. In their work, a Multi-Scale VGG16 CNN (M-VGG16) is proposed to capture more features by adding Multi-Scale Blocks (MSB) after the first four max-pooling layers of the VGG16 network. In FF-VGG16, early and late extracted features are processed independently into the MSB blocks, containing more convolutions. Instead of this method, the FF-VGG19 network proposed here does not process the extracted features further. Instead, the features are connected directly to a Global Average Pooling layer before normalising their weights by a Batch Normalisation Layer and the random disconnection of 50% of the learned weights, using a Dropout Layer. Finally, the extracted features are fused via concatenation. Moreover, early extracted features are not processed to prevent the model from learning too local and possibly insignificant information, mined from the earlier layers. The optimal parameters were defined after extensive experiments and are presented in Table 2.

3.4. *Experiment setup and methodology*

The LIDC – IDRI and CT datasets are the targets for accurate classification, achieved by employing the techniques above. The DC-GAN network is trained to generate 15,000 new and realistic lung nodules simultaneously using both datasets. Despite the independent training for benign and malignant nodules, the generated images lack reliable labelling. Therefore, a self-training algorithm is proposed to exploit the generated images. In this way, FF-VGG19 manages to exploit the most reliably labelled generated nodules to increase the training set and improve classification accuracy. The self-training algorithm is presented in Section 3.5. The new datasets created by DC-GAN are utilised for training FF-VGG19 and inspecting the performance on classifying the original datasets. During each training, data augmentation is applied to the training set. Specifically, the images are randomly rotated and flipped. In Table 3, an overview of the experimental cases is presented. An overview of the experiment is presented in Fig. 6.

Table 1 – Parameters of the components of the DC-GAN model.

Parameter	Description – Values
Generator	
Dropout	50%
Convolution's receptive field	2×2
Batch Normalisation	Across the network
Activation of the output layer	Tanh
Discriminator	
Optimiser	Adam (learning rate: 0.0003, beta1: 0.5, beta2: 0.999)
Dropout	50%
Convolution's receptive field	3×3
Batch Normalisation	No
Activation of the output layer	Sigmoid
DC-GAN	
Optimiser	Adam (learning rate: 0.0002, beta1: 0.5, beta2: 0.999)
Gaussian Noise space	500
Epochs	450
Batch Size	64

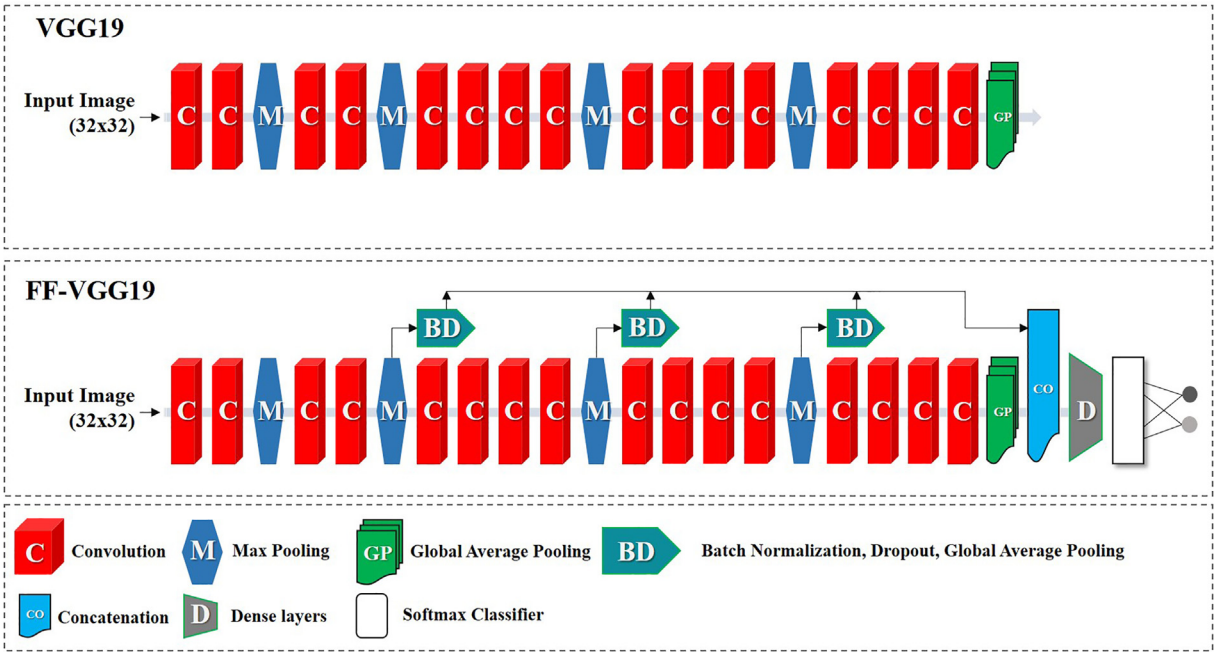

Fig. 5 – Structure of FF-VGG19.

Table 2 – The parameters of FF-VGG19.

Parameter	Values
Optimiser	Adam (learning rate: 0.0002, beta1: 0.9, beta2: 0.999)
Dropout	50%
Convolution's receptive field	3×3
Batch Normalisation	After each convolution layer
Activation of the output layer	ReLU
Neural Network at the top	2500 nodes
Classifier	Softmax
Epochs	35
Batch Size	64

Table 3 – Overview of the experimental cases and the datasets utilised. 10-fold cross-validation is used for evaluating the proposed methods. The data sizes reported referring to each of the 10 iterations. The amount of GAN-generated imaged used in training folds is not constant but may vary from 0 to 15.000, according to the efficiency of the DC-GAN and the self-training approach. The latter is performing a selection of the GAN-generated images to improve the classification performance of the classifier progressively.

Case number	Training Data	Test Data	Training Fold Size (nodules)	Test Fold Size (nodules)
1	LIDC-IDRI	LIDC-IDRI	1113	123
2	LIDC-IDRI + CT	LIDC-IDRI	1113	123
3	CT	CT	155	17
4	LIDC-IDRI + CT	CT	155	17
5	LIDC-IDRI + CT + GAN	LIDC-IDRI	GAN + 1113	123
6	LIDC-IDRI + CT + GAN	CT	GAN + 155	17
7	CT + GAN	LIDC-IDRI	1113	123
8	LIDC-IDRI + GAN	CT	155	17

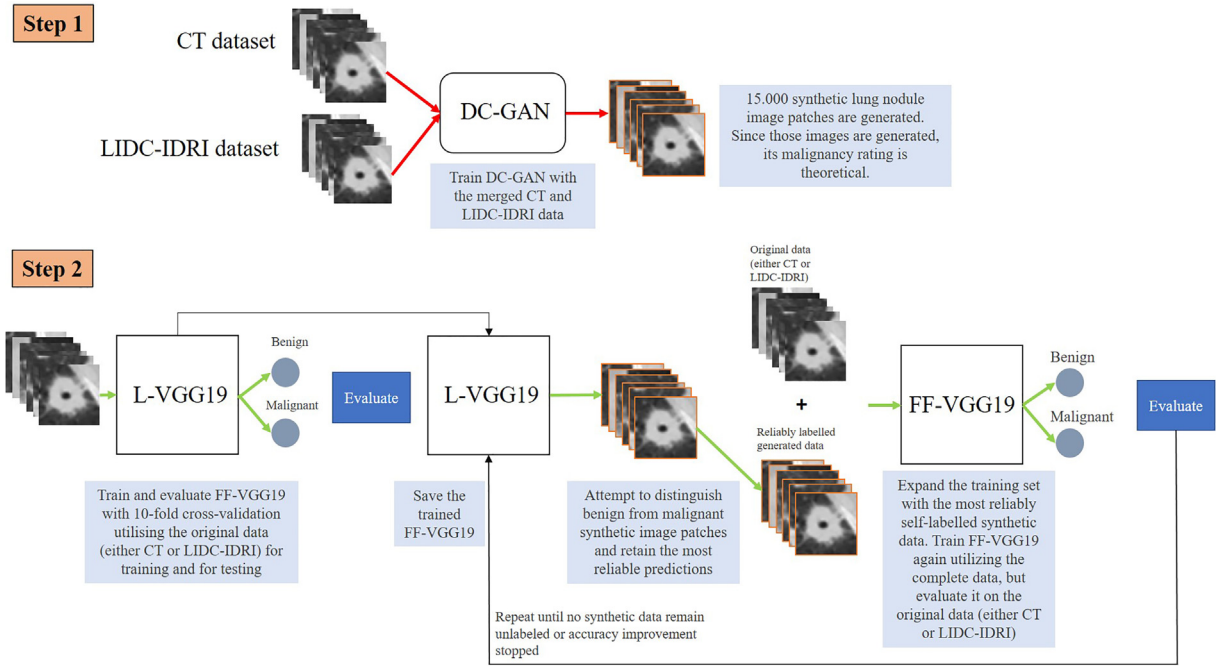

Fig. 6 – The overall procedure of the experiment.

3.5. *Self-training algorithm*

Self-training [29] is considered the simplest and one of the most efficient Semi-Supervised (SSL) algorithms [30]. This algorithm is a wrapper-based SSL approach, which constitutes an iterative procedure of self-labelling unlabeled data.

Firstly, the proposed FF-VGG19 is trained with both the LIDC-IDRI and the CT datasets. Then, those sets are iteratively augmented using CNN's most confident predictions of the unlabeled data (named GAN) generated by the DC-GAN. Next, each unlabeled image is considered reliably labelled if it has achieved a probability over a specific threshold. Finally, the reliably labelled instances are incorporated into the training data, and FF-VGG19 is retrained. This process is terminated if the following criteria are met: (a) the classification accuracy is not improving, or (b) the labelled instances are depleted. A high-level description of the Self-training algorithm applied in this experiment is presented in Table 4.

The threshold value was defined after experimenting with several values. An observation was made that the higher the threshold value, the fewer instances were picked at each iteration step, which was preferable to investigate the contribution of smaller sets over the classification accuracy.

3.6. *Evaluation procedure and metrics*

Two Nuclear Medicine experts were pooled to evaluate the nodules generated by the DCGAN and distinguish samples of fake and real nodule images. The performance of DCGAN was analysed based on the accuracy of the experts. The samples given to the experts contained a random selection of 65 real and 65 fake images.

Regarding the classification task of FF-VGG16, specific metrics were recorded as follows: (a) correctly identified malignant nodules (True Positives, TP), (b) incorrectly classified malignant nodules (False Negatives, FN), (c) correctly identified benign nodules (True Negatives, TN), and (d), incorrectly classified benign nodules (False Positives, FP). Based on those metrics, we compute the accuracy, sensitivity, and specificity of the model. In addition, we record the Area Under Curve (AUC) score based on the ROC curve, which is a graphical representation of the classifying ability for binary classifiers; we record the Area Under Curve (AUC) score.

4. Results

4.1. *Data augmentation performance of DCGAN*

Based on the evaluation strategy mentioned above, the results for the performance of DCGAN are given in Table 5. The mean accuracy of the experts (29.6%) confirms that the DCGAN manages to produce realistic nodule images potentially utilised for the classification task.

Also, a comparison between the nodule representations from the CT, LIDC-IDRI, and the GAN datasets is illustrated in Fig. 7.

Based on the generated images, some issues are observed. Firstly, the generated images lack diversity regarding the nodule shape, brightness, and position. As a result, only a few images represent a benign nodule of a well-defined shape. In addition, the nodule's surrounding tissues are seldom generated realistically. This problem was tackled by generating more nodules and selecting a group of nodules containing well-defined surroundings. These issues are related to the training dataset, i.e., the training set lacks the necessary diversity and nodules with the above characteristics are not frequent. Secondly, even though the DC-GAN was trained separately for each class, there were examples where malignant nodules were generated as benign.

Table 4 – Proposed algorithm for self-training. Several synthetic lung nodule patches are progressively characterised as benign or malignant by the classifier through this operation. After each iteration, the most reliable predictions are incorporated into the initial dataset (which contained solely authentic images in the first place). Then, the classifier is retrained utilising the expanded dataset and evaluated to improve its performance in classifying the original images (not the synthetic ones).

Information	Algorithm: Self Training
Input:	**L**: a set of labelled data (LIDC-IDRI, CT) **LS**: a set of self-labelled data (empty in the beginning) **U**: a set of unlabeled data (GAN dataset of 15,000 images) **Threshold**: 0.9995 **Classifier**: CNN (FF-VGG19)
Output:	Trained Classifier Classification Accuracies, Confusion Matrixes, AUC scores
Steps	
1:	Train-Test CNN on labelled data (L) and record the accuracy
2:	Ask CNN to predict the labels of the unlabeled samples (U)
3:	Pick reliable instances with predicted probability > threshold. The instances constitute a set called Lu
4:	Remove Lu from U, and add Lu to the set of the self-labelled data (LS). (LS = LS + Lu)
5:	Train-Test Classifier on L + LS and test on L (test folds contain only authentic images)
6:	If the classification accuracy of step 5 is better than that of step 1, continue, remove Lu from LS, and discard the instances, Lu, from U permanently.
8:	**Repeat Steps 1**...**7 until** U is empty or accuracy is not improving

Table 5 – Experts' accuracy in distinguishing real and fake nodule representations.

Expert	Fake Images Accuracy	Real Images Accuracy	Overall Accuracy
1	18 of 65	24 of 65	32,3%
2	12 of 65	23 of 65	26.9%

4.2. *Classification performance of FF-VGG19*

The classification metrics of the FF-VGG16 are presented in Table 6. The evaluation was performed with 10-fold cross-validation. In this way, the train and test split is 90–10%.

As Table 6 suggests, a classification accuracy of 92,07% of the dataset LIDC-IDRI is achieved. Besides, by utilising generated nodule images by the DC-GAN, the accuracy on the LIDC-IDRI dataset was improved by 7%. The contribution of the generated SPN images to a realistic expansion of the training dataset is also confirmed by the classification accuracy of the CT dataset, which was improved by 5%, obtaining 84,3%.

Another experiment was conducted to evaluate the strategy of labelling the generated nodules, utilising the initially generated nodules of DC-GAN. The results are illustrated in Table 7.

The classification accuracy of the SSL training scheme outperforms the classification accuracy of the case where the SPN images were incorporated into the training set, as

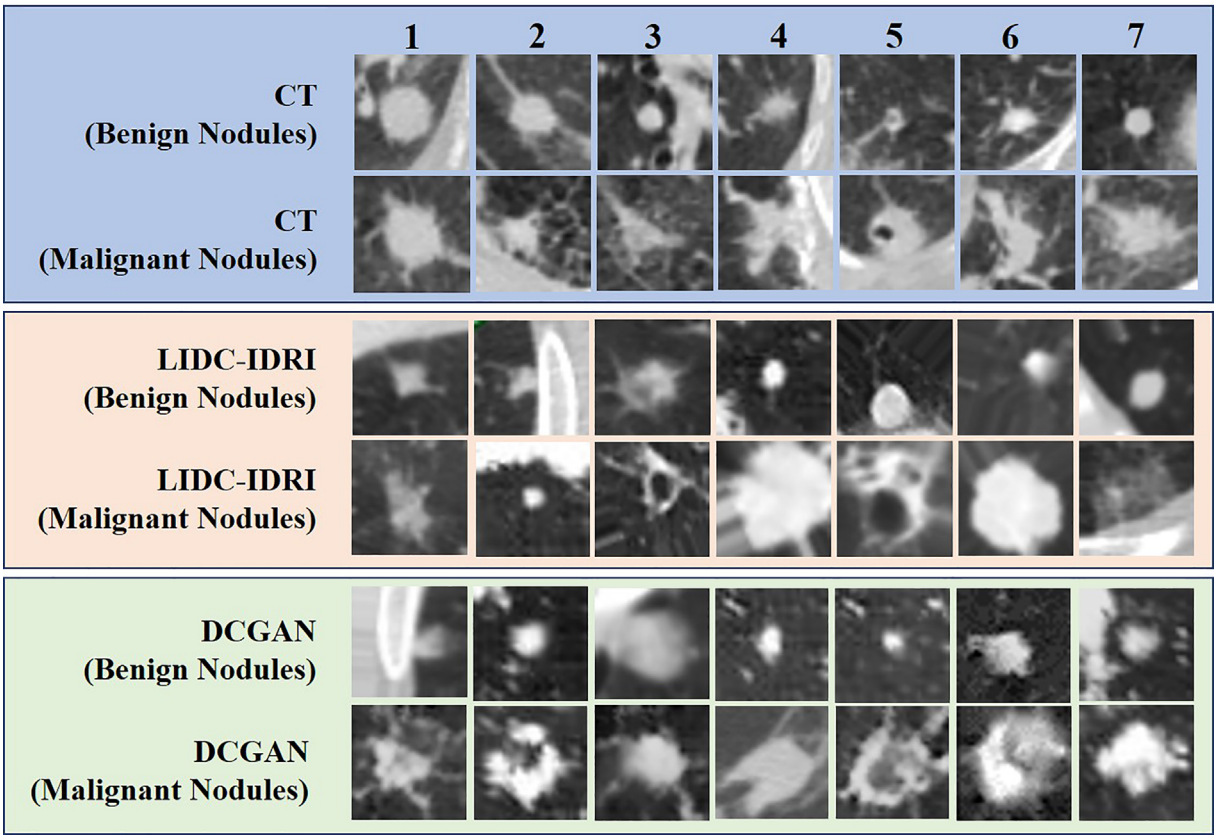

Fig. 7 – Samples from CT, LIDC-IDRI and DC-GAN SPNs of the two classes.

Table 6 – Evaluation Metrics of FF-VGG19 for the classification of the different datasets of the training process.

Case	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
1	84,95	80,16	89,77	84,96
2	85,11	83,22	87,01	85,11
3	79,06	80,45	77,64	79,05
4	80,23	77,01	83,52	80,27
5	**92,07**	**89,35**	**94,80**	**92,08**
6	84,30	82,75	85,88	84,32
7	84,46	91,12	77,75	84,44
8	80,92	80,45	81,39	80,92

labelled by the DC-GAN. The above demonstrates the advantage of labelling the generated SPNs in a semi-supervised way. However, it also highlights that the proposed DC-GAN, although trained separately for each class, cannot distinguish the characteristics between benign and malignant nodules.

Moreover, the classification performance of some alternative CNN structures was recorded. The relative accuracies of the alternative CNN structures are presented in Table 8. Specifically, for experiment 1, four BD boxes were utilised, connected with every Max Pooling layer of the VGG19. For experiment 2, two BD boxes were utilised, connected to the early extracted features (i.e., the first two Max-Pooling layers). For experiment 3, the two BD boxes were connected to the last Max-Pooling layers. For experiment 4, three BD boxes were connected to the first Max Pooling layers, while for experiment 5, the three BD boxes were connected to the last Max Pooling layers. The experiments were conducted using the training and test sets of Case 5.

4.3. *Comparisons with published results*

To further evaluate the proposed FF-VGG19, a comparison between the performance of various VGG19 and VGG16 modifications is illustrated in Table 9. VGG19_v1 refers to the architecture of VGG19, trained from scratch. VGG19_v2 refers to the architecture and the weights of FF-VGG19 without the extra BD paths. VGG16_v1 refers to VGG16 trained from scratch, while VGG16_v2 refers to the modification of VGG16 utilising the parameters mentioned in [8]. Finally, M-VGG16 refers to the modification of VGG19 proposed by Zhao et al. in [9] trained from scratch.

The layers added after the CNN structures were the same for all networks, except for the one proposed by Zhao et al., which retained the same structure, as explained in their work [9]. For the remaining CNNs, the same Neural Network structure was inserted to their tops (i.e. the final layer), which consisted of one layer of 2500 nodes, a Batch Normalisation layer and a Dropout layer to disconnect 50% of the learned weights randomly. Besides, a Global Average Pooling was applied after the Convolutional layers and before the Neural Network.

All the compared CNNs were compiled with the Adam Optimiser (learning rate of 0.001, beta1 of 0.9, and beta2 of 0.999), trained for 35 epochs, with a batch size of 64. The CNNs were trained on the dataset of Case 5 (LIDC-IDRI + CT + GAN) and evaluated with 10-fold cross-validation.

The best accuracy obtained by the proposed FF-VGG19 underlined the effectiveness of the late-feature extraction policy, where the high-level features are connected directly to the dense layers, and the early features are not further utilised.

Moreover, a comparison between the proposed classification strategy and related proposals was recorded and is presented in Tables 10 and 11. In Table 10, studies utilising solely automatic feature extraction, while their test sets are of significant size, are included. In Table 11, studies exploiting both automatic feature extraction and auxiliary attributes are mentioned. All the mentioned strategies were evaluated, utilising the LIDC-IDRI dataset. The cited studies achieved the highest classification accuracy to the best of our knowledge, utilising automatic, deep feature extraction.

The proposed methodology achieves the second-best classification accuracy. However, in the study of Wu et al. [33], the classification is based on five classes in ascending order as to their malignancy risk. At the same time, they regard an attribute/malignancy rating with +1 or −1 as an acceptable result. In this study, the classes are only benign/malignant, which means that every probability score corresponds to a single class, and no grey zone is allowed. Nevertheless, the proposed methodology achieves the best specificity score, which measures a benign nodule image's probability of being correctly predicted as benign.

The asterisk (*) indicates that the generated nodule images were incorporated into the test set or increased with augmented nodule images. The double-asterisk (**) indicates that the authors re-considered the initial labelling of the samples provided by co-operating radiologists for the LIDC-IDRI data release.

To the best of our knowledge, the related researches utilising approximately the same samples of the LIDC-IDRI dataset, and applying deep feature extraction, do not achieve above 90% classification accuracy. Hence, as the results suggest, the proposed training scheme outperforms most state-of-the-art studies, utilising large samples. Besides, the accuracy obtained by training on a small number of samples is less significant than the accuracy obtained by studies using samples approximately equal to the dataset population.

Moreover, the results highlight that studies utilising automatic deep feature extraction and handcrafted features achieve higher accuracy. This fact indicates that more research has to be conducted to improve the methods of automated feature extraction, which is, of course, a preferable strategy to reduce the workload and avoid manual feature extraction, which can be a rather painful procedure.

4.4. *Statistical significance tests*

Statistical significance tests are performed to evaluate the semi-supervised learning strategy, the effect of the BD boxes in the FF-VGG19 architecture, and the overall importance of the multi-path method of FF-VGG19.

Table 7 – Comparison of the classification performance with and without self-training.

Training Dataset	Test Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
LIDC-IDRI + CT + GAN (with SSL)	LIDC-IDRI	**92,07**	**89,35**	**94,80**	**92,08**
LIDC-IDRI + CT + GAN (without SSL)	LIDC-IDRI	81,47	74,67	88,31	81,49
LIDC-IDRI + CT + GAN (with SSL)	CT	**84,46**	91,12	77,75	84,44
LIDC-IDRI + CT + GAN (without SSL)	CT	79,11	79,51	78,66	79,09

Table 8 – Classification performance of FF-VGG19, utilising different BD boxes.

Experiment	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
1	88,75	88,23	89,29	88,76
2	89,07	88,55	89,61	89,08
3	91,74	86,61	96,92	91,76
4	87,78	87,42	88,15	87,78
5	**92,07**	**89,35**	**94,80**	92,08

For the statistical significance tests, the procedure suggested by Dietterich [40] is followed. 10 times 2-fold cross-validation is performed to ensure that the same image will not be part of the train and test sets more than once under the k-fold cross-validation procedure. First, the mean fold accuracy of each of the ten iterations is recorded. Next, the mean accuracy difference *t*-test for the two samples is performed to investigate the statistical significance of the mean accuracy between the compared experiments (Table 12).

The statistical significance of the mean accuracy of the semi-supervised and supervised learning strategy in LIDC-IDRI dataset classification is observed. P-value is found to be less than 0.05, and T-statistic is found to be 7.56. In CT, the p-value is 0.001 and T-statistic 4.05, which implies statistical significance. Regarding the number of paths (BD boxes) of FF-VGG19, the results show no statistical significance in the accuracy between the two compared experiments. Therefore, FF-VGG19 operates with 4 or 5 paths, yielding approximately the same results. Finally, comparing the best performing VGG19 (v2) and FF-VGG19 yields a p-value of 0.09 and a T-statistic of −1.48. It is concluded that the mean accuracy between the two networks is statistically significant. The results confirm the effectiveness of the semi-supervised learning strategy and the superiority of FF-VGG19 over the traditional VGG19, at least for the medical classification task under investigation.

5. Discussion

The results of this study highlight the effectiveness of GANs, especially the network structure of DC-GAN, in generating new, realistic lung nodule images. The evaluation criteria were two-fold. Firstly, medical experts did not easily distinguish the fake samples, demonstrating their similarity with actual nodule samples. Secondly, the generated samples benefited the performance of the FF-VGG19, which also proves their significance. Moreover, supplying the discriminator of the DC-GAN with nodule images from both LIDC-IDRI and CT datasets catalysed the generation of more realistic and diverse nodules.

As far as the classification task is concerned, retaining information from late-extracted features from VGG19 by connecting each late convolution group into an independent path achieved better results than other VGG16 and VGG19 architectures may suggest that the early extracted features are indeed less significant. Moreover, this indicates that the deeper the network is during feature extraction, the more information is fuzzed or lost unless a multi-path structure is utilised.

As observed and proved by the results, despite the similarity of the fake images with the real ones, there was a problematic categorisation by the DCGAN of the classes they belong to, whether they are benign or malignant. This issue could be overcome by implementing more sophisticated GAN networks, such as Info-GAN [39]. However, human interaction was necessary to remove some unrealistic nodule image patches, which is inevitable in such methods. Furthermore, during the generation of 15,000 images, few images are always expected to be inappropriate. Therefore, further attempts and research should be directed towards improving the generation capabilities of GANs to achieve the automatic discard of unrealistic images.

Even though the self-training algorithm of this study is proven to improve the classification performance of CNN, it is not necessarily the optimal strategy. Therefore, more approaches to semi-supervised learning could be investigated and compared.

Nevertheless, this study presents an effective strategy for the accurate classification of lung nodules from CT scans, achieving one of the best accuracies, sensitivity, and specificity compared to the state-of-the-art. Moreover, it has been proven that generating new training samples utilising DC-GAN is a successful approach to expand the training set to

Table 9 – Performance comparison between alternative VGG structures trained on the dataset of Case 5.

CNN	LIDC-IDRI			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
VGG19_v1	89,48	84,19	94,81	89,50
VGG19_v2	90,04	87,74	92,37	90,06
VGG16_v1	87,94	90,00	85,88	87,94
VGG16_v2	88,34	90,48	86,20	88,34
M – VGG16 [9]	90,37	88,71	92,05	90,38
FF-VGG19	**92,07**	89,35	94,81	92,08

Table 10 – Comparison with studies utilising only automatic feature extraction, while their test sets are of significant size.

Author	Strategy	Data selection methodology	Size of the test set	ACC (%)	SEN (%)	SPE (%)	AUC (%)
Cheng [32]	Stacked Autoencoder (SAE)	Nodules with a mean malignancy rating in [1,2] were selected as benign. Nodules with a mean malignancy rating in [4,5] were selected as malignant. The rest are excluded.	1400	87,4	86,3	88,5	94,1
Wu [33]	CNN	Nodules with a mean malignancy rating of 3 are excluded. The rest are separated into two classes with a threshold of 3.	1404	**97,6**	–	–	–
Dey [34]	Transfer Learning	Nodules diagnosed by at least three radiologists are considered. The median value is used to define the overall rating. Nodules with a median value above 3 are considered malignant. Nodules with a median value equal to 3 are excluded.	686	90,4	90,4	90,3	**95,5**
Shen [35]	CNN	Mean value is used to define the overall malignancy rating. Nodules with a mean rating value above 3 are considered malignant. Nodules with a mean rating value equal to 3 are excluded.	1375	87,1	77,0	93,0	93,0
This study	GAN + Transfer Learning	Nodules with mean malignancy rating in the interval [2.5, 3.5] are excluded.	1236	92,1	89,3	**94,8**	92,08

Table 11 – Comparison with studies exploiting both automatic feature extraction and auxiliary attributes is mentioned or augmenting the test sets.

Author	Strategy	Data selection methodology	Size of the test set	ACC (%)	SEN (%)	SPE (%)	AUC (%)
Zhao et al. [9]	GAN + Transfer Learning	Nodules with a mean malignancy rating in [1,2] were selected as benign. Nodules with a mean malignancy rating in [4,5] were selected as malignant. The rest are excluded.	353	**95,2**	**98,6**	92,4	98,0
Yang et al. [22]	CNN	Nodules with a majority score ≥ 4 are considered malignant, and the rest benign.	2025**	91,7	58,3	97,7	88,2
Cheng [32]	Stacked Autoencoder (SAE)	Nodules with a mean malignancy rating in [1,2] were selected as benign. Nodules with a mean malignancy rating in [4,5] were selected as malignant.	10,133*	94,4	90,8	98,1	**98,4**
Song et al. [36]	CNN	Nodules with a mean malignancy rating in [1,2] were selected as benign. Nodules with a mean malignancy rating in [3,4,5] were selected as malignant.	5024*	84,1	83,9	84,3	91,6
Causey [37]	CNN + Radiomics	Nodules with a mean malignancy rating in [1,2] were selected as benign. Nodules with a mean malignancy rating in [4,5] were selected as malignant. The rest are excluded.	664	93,2	87,9	**98,5**	97,1
Da Silva [38]	CNN + Genetic Algorithm	Nodules with a mean malignancy rating in [1,2] were selected as benign. Nodules with a mean malignancy rating in [3,4,5] were selected as malignant. The rest are excluded.	1343*	94,8	94,7	95,1	94,9
This study	GAN + Transfer Learning	Nodules with mean malignancy rating in the interval [2.5, 3.5] are excluded.	1236	92,1	89,3	94,8	92,1

Table 12 – Statistical significance test results.

	Test Dataset	Mean Accuracy (%)	Standard Deviation (%)	Degrees of Freedom	T-statistic	p-value
SSL (see Table 7) No-SSL (see Table 7)	LIDC-IDRI LIDC-IDRI	88.30 75.89	2.7 3.87	9	7.56	0.00003
SSL (see Table 7) No-SSL (see Table 7)	CT CT	79.12 72.96	2.26 3.22	9	4.05	0.001
FF-VGG19 with 4 BD boxes FF-VGG19 with 5 BD boxes	LIDC-IDRI LIDC-IDRI	86.69 88.30	2.2 2.7	9	−1.22	0.12
VGG19_v2 FF-VGG19	LIDC-IDRI LIDC-IDRI	86.51 88.30	2.47 2.7	9	−1.48	0.09

train robust CNNs. Finally, the proposed FF-VGG19, trained with the semi-supervised learning method, circumvents the labelling issue, outperforming several related models and strategies.

6. Conclusions

The application of GANs for problems involving the processing of medical images is a high-risk choice due to the importance of the biomarkers acquired from medical images; still, it is an option that can help in more comprehensive training of diagnostic models. The targeted creation of unreal but realistic nodule representations, which were utilised to expand the training set, improved the efficiency of CNN for the characterisation of the authentic nodules' malignancy rating. The above suggests that DC-GANs can be efficiently employed to solve medical image tasks, where the available data are causing non-trivial issues in CNN training.

In future work, new methods of utilising the weakly labelled generated SPN images with semi-supervised training schemes will be investigated to ensure that accurate labelling and improved classifier performance are gradually achieved. More specifically, the intention is to experiment with self-training, co-training, and tri-training while also employing CST-Voting. Livieris et al. [31] proposed combining the three methods and perhaps obtaining better results.

Funding

No funding was received for conducting this study.

Data availability statement

The LIDC-IDRI dataset analysed during the current study is available in the Cancer Imaging Archive repository https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.

The CT dataset analysed during the current study is not available due to ethical reasons.

Code availability

The complete code is available in Github: https://github.com/apjohndim/SPN-Classification-with-GANs-MVGG19.

CRediT authorship contribution statement

Ioannis D. Apostolopoulos: Conceptualization, Software, Writing - original draft, Visualization, Validation. **Nikolaos D. Papathanasiou:** Methodology, Writing - review & editing, Validation. **George S. Panayiotakis:** Methodology, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Postmus P, Kerr K, Oudkerk M, Senan S, Waller D, Vansteenkiste J, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2017;28: iv1–21.

[2] Zhuang H, Pourdehnad M, Lambright ES, Yamamoto AJ, Lanuti M, Li P, et al. Dual time point 18F-FDG PET imaging for differentiating malignant from inflammatory processes. J Nucl Med 2001;42:1412–7.

[3] Tajbakhsh N, Suzuki K. Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: MTANNs vs CNNs. Pattern Recogn 2017;63:476–86.

[4] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[5] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in neural information processing systems 27. Curran Associates, Inc.; 2014. p. 2672–80.

[6] Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans: The LIDC/IDRI thoracic CT database of lung nodules. Med Phys 2011;38:915–31. https://doi.org/10.1118/1.3528204.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:14091556 [cs] 2015.

[8] Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilising transfer learning with

convolutional neural networks. Phys Eng Sci Med 2020;43:635–40. https://doi.org/10.1007/s13246-020-00865-4.
[9] Zhao D, Zhu D, Lu J, Luo Y, Zhang G. Synthetic medical images using F&BGAN for improved lung nodules classification by multi-scale VGG16. Symmetry 2018;10:519.
[10] Chuquicusma MJ, Hussein S, Burt J, Bagci U. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE; 2018, p. 240–4.
[11] Javaid U, Lee JA. Capturing variabilities from computed tomography images with generative adversarial networks. arXiv:180511504 [cs, stat] 2018.
[12] Jin D, Xu Z, Tang Y, Harrison AP, Mollura DJ. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. International conference on medical image computing and computer-assisted intervention, Springer; 2018, p. 732–40.
[13] Harrison AP, Xu Z, George K, Lu L, Summers RM, Mollura DJ. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. International conference on medical image computing and computer-assisted intervention, Springer; 2017, p. 621–9.
[14] Han C, Kitamura Y, Kudo A, Ichinose A, Rundo L, Furukawa Y, et al. Synthesising diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection. arXiv:190604962 [cs, eess] 2019.
[15] Tang Y, Cai J, Lu L, Harrison AP, Yan K, Xiao J, et al. CT image enhancement using stacked generative adversarial networks and transfer learning for lesion segmentation improvement. arXiv:180707144 [cs] 2018.
[16] Rother C, Kolmogorov V, Blake A. "GrabCut" interactive foreground extraction using iterated graph cuts. ACM transactions on graphics (TOG) 2004;23:309–14.
[17] Xie S, Tu Z. Holistically-nested edge detection. Proceedings of the IEEE international conference on computer vision, 2015, p. 1395–403.
[18] Yan K, Wang X, Lu L, Zhang L, Harrison AP, Bagheri M, et al. Deep lesion graphs in the wild: relationship learning and organisation of significant radiology image findings in a diverse large-scale lesion database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, p. 9261–70.
[19] Bi L, Kim J, Kumar A, Feng D, Fulham M. Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs). In: Cardoso MJ, Arbel T, Gao F, Kainz B, van Walsum T, Shi K, editors. Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment. Cham: Springer International Publishing; 2017. p. 43–51. https://doi.org/10.1007/978-3-319-67564-0_5.
[20] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Ann Hist Comput 2017:640–51.
[21] Onishi Y, Teramoto A, Tsujimoto M, Tsukamoto T, Saito K, Toyama H, et al. Automated pulmonary nodule classification in computed tomography images using a deep convolutional neural network trained by generative adversarial networks. BioMed Res Int 2019;2019:1–9. https://doi.org/10.1155/2019/6051939.
[22] Yang J, Liu S, Grbic S, Setio AAA, Xu Z, Gibson E, et al. Class-aware adversarial lung nodule synthesis in CT images. 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), IEEE; 2019, p. 1348–52.
[23] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57. https://doi.org/10.1007/s10278-013-9622-7.
[24] Shi W, Caballero J, Theis L, Huszar F, Aitken A, Ledig C, et al. Is the deconvolution layer the same as a convolutional layer? arXiv preprint arXiv:160907009 2016.
[25] Ioffe S, Szegedy C. Batch normalisation: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167 2015.
[26] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
[27] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434 2015.
[28] Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:150500853 2015.
[29] Kingma DP, Ba J. Adam: A method for stochastic optimisation. arXiv preprint arXiv:14126980 2014.
[30] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. In: 33rd annual meeting of the association for computational linguistics. p. 189–96.
[31] Livieris IE. A new ensemble self-labeled semi-supervised algorithm. Informatica 2019;43.
[32] Cheng J-Z, Ni D, Chou Y-H, Qin J, Tiu C-M, Chang Y-C, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2016;6:1–13.
[33] Wu B, Zhou Z, Wang J, Wang Y. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. arXiv:180203584 [cs] 2018.
[34] Dey R, Lu Z, Hong Y. Diagnostic classification of lung nodules using 3D neural networks. arXiv preprint arXiv:180307192 2018.
[35] Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, et al. Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification. Pattern Recogn 2017;61:663–73. https://doi.org/10.1016/j.patcog.2016.05.029.
[36] Song Q, Zhao L, Luo X, Dou X. Using deep learning for classification of lung nodules on computed tomography images. J Healthc Eng 2017;2017.
[37] Causey JL, Zhang J, Ma S, Jiang B, Qualls JA, Politte DG, et al. Highly accurate model for prediction of lung nodule malignancy with CT scans. Sci Rep 2018;8:9286.
[38] da Silva GL, da Silva Neto OP, Silva AC, de Paiva AC, Gattass M. Lung nodules diagnosis based on evolutionary convolutional neural network. Multimedia Tools Appl 2017;76:19039–55.
[39] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. Infogan: Interpretable representation learning by information maximising generative adversarial nets. Adv Neural Inf Process Syst 2016:2172–80.
[40] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10:1895–923. https://doi.org/10.1162/089976698300017197.