Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-021-02378-y

ORIGINAL ARTICLE

Automatic classification of solitary pulmonary nodules in PET/CT imaging employing transfer learning techniques

Ioannis D. Apostolopoulos[1] · Emmanuel G. Pintelas[2] · Ioannis E. Livieris[2] · Dimitris J. Apostolopoulos[3] · Nikolaos D. Papathanasiou[3] · Panagiotis E. Pintelas[2] · George S. Panayiotakis[1]

Received: 21 March 2020 / Accepted: 6 May 2021

Abstract

Early and automatic diagnosis of Solitary Pulmonary Nodules (SPN) in Computed Tomography (CT) chest scans can provide early treatment for patients with lung cancer, as well as doctor liberation from time-consuming procedures. The purpose of this study is the automatic and reliable characterization of SPNs in CT scans extracted from Positron Emission Tomography and Computer Tomography (PET/CT) system. To achieve the aforementioned task, Deep Learning with Convolutional Neural Networks (CNN) is applied. The strategy of training specific CNN architectures from scratch and the strategy of transfer learning, by utilizing state-of-the-art pre-trained CNNs, are compared and evaluated. To enhance the training sets, data augmentation is performed. The publicly available database of CT scans, named as Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI), is also utilized to further expand the training set and is added to the PET/CT dataset. The results highlight the effectiveness of transfer learning and data augmentation for the classification task of small datasets. The best accuracy obtained on the PET/CT dataset reached 94%, utilizing a modification proposal of a state-of-the-art CNN, called VGG16, and enhancing the training set with LIDC-IDRI dataset. Besides, the proposed modification outperforms in terms of sensitivity several similar researches, which exploit the benefits of transfer learning.

Keywords Solitary pulmonary nodule classification · Deep learning · Convolutional Neural Networks · Transfer learning · Data augmentation

1 Introduction

Low-dose lung Computer Tomography (CT) screening provides an effective way for the early diagnosis of lung cancer. Commonly, clinicians observe, analyze, and interpret the CT scans according to the results of nodule morphology and clinical conditions [1]. Due to the human's physical factors, such as limitation of the visual system, fatigue, and distraction, clinicians may not make the best use of the CT image data [2]. Moreover, clinicians may have to analyze a vast number of scans daily, which is a time-consuming procedure.

An automatic analysis of medical images, using computer aided diagnostic systems could make a significant contribution not only of workload reduction, but also in the early treatment of cancer. Despite the significant progress, fully automatic detection and characterization of lung nodules is still an open issue, due to a variety of reasons, such as the absence of reliably labeled, large-scale datasets [3]. The incorporation of Machine Learning (ML) and Deep Learning (DL) techniques for classification and segmentation tasks on medical images is a new promising technique for the last decade [4].

Deep Learning, and more specifically, Convolutional Neural Networks (CNN) [5], alludes to a wide class of Machine Learning methods and structures, utilizing large amount of data. Therefore, the networks' architectures consist of many processing layers, thereby learning both simple and complex information [6].

✉ Ioannis D. Apostolopoulos
ece7216@upnet.gr

[1] Department of Medical Physics, School of Medicine, University of Patras, 26504 Patras, Greece

[2] Department of Mathematics, University of Patras, 26504 Patras, Greece

[3] Laboratory of Nuclear Medicine, University of Patras, 26504 Patras, Greece

Published online: 18 May 2021

CNNs have proven to be powerful tools for a broad range of computer vision tasks. Since 2010, the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [7] has brought dramatic progress in image processing. This competition is based on the ImageNet database, which contains over 14 million images belonging to 1000 classes and annotated by the human hand. Many CNNs have been proposed over the last years to classify those images, such as the CNN created by and named after the Visual Geometry Group (VGG) of the University of Oxford [8], and the CNN proposed by Howard et al. [9], called MobileNet.

Before the advances in Deep Learning, manual feature engineering followed by classifiers was the general pipeline for abnormally detection, with remarkable results [10]. However, manual feature extraction is still a time-consuming procedure, while there is a dispute between the researchers as to which feature is significant and which is not. After the relatively large-scale Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [11] dataset became publicly available, deep learning-based methods have become the dominant framework for nodule classification research [12].

For this study, the problem of automatic characterization of 167 single Solitary Pulmonary Nodules (SPNs) from a small dataset of CT images, extracted from 112 Positron Emission Tomography and Computer Tomography (PET/CT) scans at the Laboratory of Nuclear Medicine of the University of Patras, is considered. The dataset consisting of those nodules is referred to as PET/CT dataset.

For the classification task, and especially when dealing with small datasets, two strategies related to the definition of training schemes and CNN structures are commonly used. The first strategy involves developing experimental CNN structures, or building on existing and reliable state-of-the-art CNNs, by exploiting parts of their architecture, and then training the CNNs from scratch [13]. The second strategy, which is a mean of what is called transfer learning, retains both the main body and specific learned weights of a pre-trained CNN, while it performs certain modifications for optimal results. In this study, both strategies are applied and compared to obtain the best result.

Since the PET/CT dataset, which is the target of accurate classification, contains a small number of unique nodules to train a deep and robust CNN, data augmentation is preferred to increase the training set with generated nodule images.

To further increase the amount of training data, it is proposed to join LIDC-IDRI and PET/CT dataset and to evaluate the performance of the employed networks. Although LIDC-IDRI and PET/CT datasets consist of images of the same nature (i.e., CT scans), their characteristics may vary, due to technical differences between the two CT scanners, or the fact that CT scans conducted with PET/CT were acquired in free-breathing conditions. For this reason, the CNNs have to be trained to ignore such diverse characteristics (e.g., texture, nodule position, contrast). The combination of PET/CT and LIDC-IDRI nodule representations resulted in enhanced robustness of the CNNs.

The contribution of this research lies in demonstrating that transfer learning can be an effective strategy in extracting the representative imaging biomarkers from chest CT images extracted from a PET/CT images, and that it is also a robust and preferable strategy to circumvent the shortage of large-scale datasets to train deep and effective networks. The experimental results prove the effectiveness of transfer learning by using pre-trained networks, due to the fact that their accuracy outperformed the performance of experimental CNN architectures, which were trained from scratch. Moreover, the proposed transfer learning scheme obtains the highest sensitivity, which is a measure in defining the true positive rate of a test, compared to similar researches.

The remainder of the paper is divided into the following sections: Section 2 illustrates recent related work. In Section 3, a brief description of transfer learning and data augmentation is provided. The research methodology of the experiment and the methods utilized are also illustrated in Section 3. The results are shown in Section 4. Discussion and concluding remarks are presented in Sections 5 and 6, accordingly.

2 Related work

The problem of nodule classification, which is also considered as a problem of lung suspicious classification of lung nodule malignancy, is considered in a variety of studies. Due to the shortage of large-scale medical image datasets to train state-of-the-art networks presented in Image Net challenge from scratch, the transfer learning approach was proposed [14]. In this section, specific studies utilizing either deep feature extraction with experimental CNNs, or transfer learning with pre-trained CNNs, are highlighted.

Zhao and Liu [15] proposed an architecture called Agile CNNs for lung nodule classification. A hybrid CNN of LeNet [16] and AlexNet [17] is constructed by combining the layer settings of LeNet and the parameter settings of AlexNet. Through adjusting the parameters of the kernel size, learning rate, and other factors, the effect of these parameters on the performance of the CNN model was investigated and an optimized setting was obtained. The model was trained and tested with 743 CT images (368 benign and 375 malignant nodules) from LIDC-IDRI. It achieved an accuracy of 82.2% and AUC of 0.877. The results have shown that the proposed CNN framework and the optimization strategy for the CNN parameters might be suitable for pulmonary nodule classification when characterized by small medical datasets and small targets.

Wei Shen et al. [18] presented a Multi-Crop Convolutional Neural Network (MC-CNN) to automatically extract salient nodule information by employing a novel multi-crop pooling strategy which crops different regions from convolutional feature maps and then applies max-pooling different times. The experiments were conducted utilizing LIDC-IDRI dataset and by augmenting the dataset using rotation, image translation, and flip. Extensive experimental results showed that the proposed method achieved 87.14% nodule suspiciousness classification performance and also characterized nodule semantic attributes (subtlety and margin) and nodule diameter which are potentially helpful in modeling nodule malignancy.

Song et al. [19] proposed two types of Artificial Neural Networks, named as Deep Neural Network (DNN) and Stacked Auto – Encoder (SAE), and one CNN for the classification task. The networks were tested using the LIDC-IDRI dataset. The extracted nodules were modified. The image of the pulmonary nodules was obtained by binary processing, which obtains the approximate outline of the pulmonary nodules. Then, the value of the pulmonary nodules was restored in the proceeded image to the pixels of the pulmonary nodules. Finally, noise disturbance around the pulmonary nodules was eliminated. The experimental results suggest that the CNN archived better performance (84.15%) than the DNN (82.37%) and SAE (82.59%).

Causey et al. proposed a systematic approach to predict lung nodule malignancy from CT data in [20]. The proposed model, named as NoduleX, was developed with a deep CNN architecture, capable of performing classification or producing a feature vector that can be used as input to a secondary classifier. The model can integrate deep learning CNN features (CNN feature expression) with radiological quantitative image features (radiomic expression) if the segmentation of the nodules is available. Utilizing a selection of reliably labeled 664 nodules from the LIDC-IDRI dataset, the model achieves 93.2% accuracy with a single partition evaluation.

Dey et al. [21] developed four two-pathway CNNs, including a basic 3D CNN, a novel multi-output network, a 3D Dense Net [22], and an augmented 3D Dense Net with multi-outputs. The CNNs are evaluated on 686 nodule images from the LIDC-IDRI dataset, obtaining 90.4% accuracy and 95.4% AUC score. In addition, the networks pretrained on the LIDC-IDRI dataset were utilized to handle a smaller dataset of 147 CT scans using transfer learning, obtaining 86.84% accuracy and 90.10% AUC score.

De Nobrega et al. [23] explored the performance of deep transfer learning for lung nodules malignancy classification. For this purpose, state-of-the-art CNNs were used as feature extractors to process the LIDC-IDRI dataset. In their work they also compared different classifiers' performance in classifying the extracted features. They concluded that the best combination of feature extractor and classifier was CNN-ResNet50 with the SVM-RBF classifier. This combination achieved 88.41% accuracy and Area Under Curve (AUC) score of 93.19%. They demonstrated that pretrained networks achieved equivalent accuracy, compared to networks designed and trained specifically for the same task, even though the pretrained networks were designed for and trained on non-medical images.

Zhao et al. [24] implemented three strategies to classify malignant and benign pulmonary nodules using CT images from LIDC-IDRI dataset, by making use of state-of-the-art CNNs. The first strategy was to perform specific modifications to the pretrained networks to specialize for the classification task of the study. The modified networks were trained from scratch. The second strategy was the integration of different network architectures and parameters and inspection of their performance. The third strategy was to employ transfer learning and fine-tuning. For the latter, four state-of-the-art models were employed and fine-tuned. In total, eleven deep CNN models were compared using the same dataset. The experimental results demonstrated that the transfer learning with fine-tuning outperforms the other two CNN schemes, which train the CNN structures from scratch.

Xie et al. [25] also adopted transfer learning for lung nodule classification. The algorithm proposed in their work transfers the image representation abilities of three ResNet-50 models. Then, the extracted features were utilized to classify lung nodules with an adaptive weighting scheme learned during the error backpropagation. The final results were obtained by weighting these models. The algorithm showed a classification accuracy of 93.40% when tested on 1357 nodules from the LIDC-IDRI dataset.

A significant limitation of the mentioned related researches is the fact that the extracted image-features are not further evaluated as to their validity and to their relation to the common characteristics of a pulmonary nodule (e.g. sphericity, calcification). In fact, it is not known whether the deep learning algorithms based their predictions on global characteristics or local features, which may be present only in the particular image dataset and do not constitute reliable indicators. Furthermore, one should notice that CNNs, as all deep learning models being "Black Box" algorithms, lack the ability to support Interpretability and Explainability of their predictions, as has been demonstrated in [26, 27].

While the LIDC-IDRI dataset has proven to be large and diverse enough to employ both transfer learning techniques and training from scratch, in this work, the performance of transfer learning is applied to a very small dataset, lacking the diversity of information incorporated into a large-scale lung nodule dataset. Besides, it is yet to be proven whether the extracted features from the LIDC-IDRI dataset are indeed significant features related to lung nodule malignancy. The latter issue can be addressed by utilizing both the LIDC-IDRI dataset and an alternative dataset for test set, which is used to evaluate the significance of the extracted features.

3 Methods

In the current section, detailed methodology of the research is provided. In Section 3.1, the fundamental questions of this work are presented. Besides, an overview of the experiments is given. In Section 3.2, the advantages of Transfer Learning and Data Augmentation are presented. The datasets of the experiment and the augmentation methods are explained in Section 3.3. The CNNs utilized for the classification of the image datasets are presented in Section 3.4.

3.1 Research methodology

The basic research questions of this work can be summarized as:

- Transfer learning to be an effective strategy for extracting representative imaging biomarkers from chest CT images and,
- To be employed as a robust and preferable strategy to circumvent the shortage of large-scale datasets to train deep and effective networks

By extracting CT images from a PET/CT system (General Electric Healthcare: Discovery iQ3 sl16), a small lung nodule image dataset emerged and is extensively described in Sect. 3.3. Due to limitations of the training size, LIDC-IDRI dataset was exploited to enhance the size of the training sets.

For the classification process, CNN architectures developed by the authors to train from scratch and state-of-the-art CNNs for transfer learning were utilized. The strategy of transfer learning was evaluated by employing a VGG network with 16 weight layers, called VGG16 [8], and another CNN, called MobileNet [9]. The selection of the specific CNNs was based on their capabilities of processing images of 32 × 32 pixel size, which corresponds to the size of the images belonging to the PET/CT dataset. The strategy of training from scratch was evaluated by developing three experimental architectures of CNNs, designed to perform feature extraction and classification, by employing different approaches, such as early feature extraction or shallow-type feature extraction.

Since the aim of the experiments is to achieve an accurate classification of the PET/CT dataset, the accuracy of the networks was measured based on the classification of PET/CT images and regardless of the training dataset. An overview of the experiment is illustrated in Fig. 1.

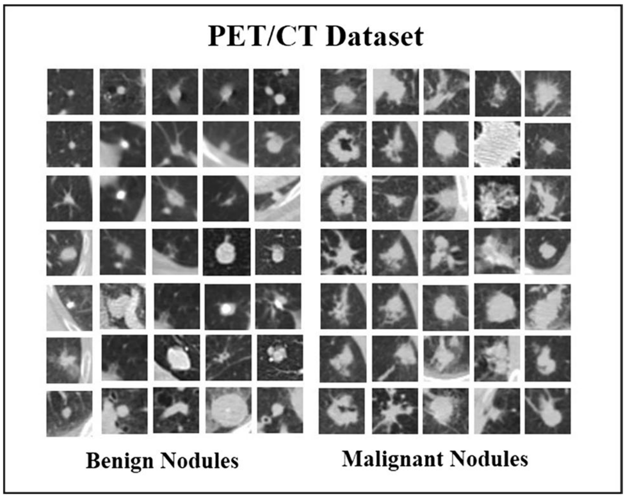

Fig. 1 Overview of the experiment's process

3.2 Deep learning with transfer learning and data augmentation

While the effectiveness of deep learning depends on the availability of data, seldom are the available datasets large enough to train robust networks. To overcome this issue, data augmentation techniques have been proposed to expand the training data. Besides, transfer learning is an established method, which can be employed especially when the available datasets do not meet the requirements of deep networks designed to be trained from scratch [16].

3.2.1 Transfer learning

Transfer learning refers to an approach wherein the knowledge (i.e. the leaned weights of each feature) mined by a learning model to a specific task is applied to solve a different task [16]. The process of transfer learning in Deep Learning involves the initial training of a CNN in one task, utilizing large datasets. The CNN structure and parameters must be equipped with the ability to generalize and learn global features, rather than sticking to local characteristics, such as pixel brightness values in a specific area of the image [12]. The pre-trained network may then be used to make predictions on a new set of images of another nature, without modifications regarding the architecture or the learned weights. Employing the pre-trained model without modifications and simply training a new network on top of it, is commonly called feature extraction via transfer learning [28]. The pre-trained model is only used as a feature extractor; the extracted features are then inserted into a new network that performs the classification task. This method is commonly used either to circumvent computational costs coming with training a very deep network from scratch, or to retain the useful feature extractors trained during the initial stage.

A common practice to perform transfer learning involves utilizing CNNs participated and stood out in the ILSVRC challenge.

3.2.2 Data augmentation

Data augmentation is an essential technique that enhances the training set of a network and is used mainly when the training dataset contains only a few samples [29].

Geometric distortions or deformations are often utilized to either increase the number of samples for deep network training, or to balance the size of datasets. In the case of microscopical images, shift and rotation invariance, as well as robustness for deformations and gray value variations are the necessary alterations applied to each image of the training set [30].

These methods have been proven fast, reproducible and reliable [18]. Increasing the number of the data may effectively improve the CNN's training and testing accuracy, reduce the loss, and improve the network's robustness. In the research for lung nodule detection, segmentation and classification, data augmentation techniques have been employed recently [31]. However, heavy data augmentation should be carefully considered, as this may produce unrealistic images and confuse the CNN.

3.3 Datasets and data augmentation

In this section, the datasets used for the experiments and the data pre-processing steps for preparation of the images are described.

4 PET/CT dataset

The dataset named as PET/CT dataset included 112 CT scans corresponding to 112 patients. For every scan, more than one SPN may be present. From those CT scans, 80 unique benign nodules and 87 unique malignant nodules were examined. For every nodule, one to seven slices that represented various aspects of the nodule was extracted.

Nodules larger than 30 mm are considered malignant and no automatic characterization is needed. Besides, nodules smaller than 3 mm are not considered Solitary Pulmonary Nodules. Therefore, nodules larger than 30 mm and smaller than 3 mm were excluded from the dataset. All SPNs meeting the above criteria were extracted.

The final PET/CT dataset consisted of 545 benign and 607 malignant nodule images. The nodule images were extracted in tiff format at a fixed size of 32 × 32 pixels. In this way, the retention of the nodule's characteristics was achieved. An area of interest of 32 × 32 pixels is enough for a nodule of 30 mm to fit in. The nodules were placed in the center of the picture. The reduction of the area of interest to 32 × 32 pixels was supervised by a Nuclear Medicine expert, to ensure that significant neighboring information was not excluded.

Labeling had been done by the physicians using (a) biopsy results, (b) Fluoro-Deoxy-Glucose (FDG) uptake, or (c) patient follow up. Weakly labeled instances, i.e., with uncertainty as to their malignancy, were excluded from the dataset.

The complete generation of the PET/CT dataset from the initial CT Scans is illustrated in Fig. 2.

5 LIDC-IDRI dataset

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) consists of 1018 diagnostic and lung cancer screening chest CT scans with marked-up annotated lesions by four radiologists. This dataset was initiated by the National Cancer Institute (NCI) [32]. For nodule extraction and crop, the annotations provided by the database were used. As in the PET/CT dataset, nodules larger than 30 mm and smaller than 3 mm were excluded. Moreover,

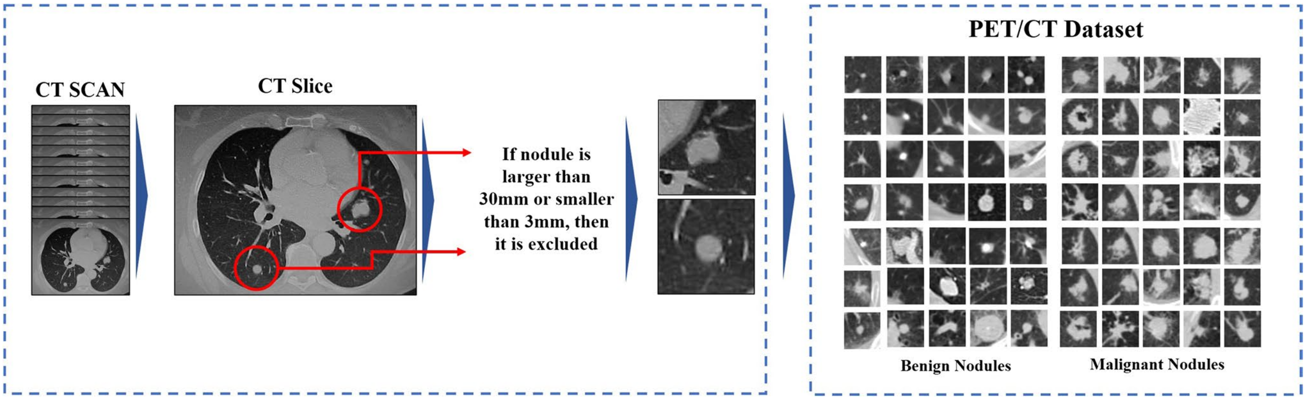
CT SCAN
CT Slice
If nodule is larger than 30mm or smaller than 3mm, then it is excluded
PET/CT Dataset
Benign Nodules
Malignant Nodules

Fig. 2 Random extracted SPN images, from the PET/CT dataset

weakly labeled nodules were excluded as their malignancy rating is not scientifically validated. For every scan, more than one SPN may be present. All SPNs meeting the above criteria were extracted.

Since the specific dataset is utilized for data augmentation purposes, noisy or low-resolution nodule representations were not incorporated. This process resulted in the selection of 1116 single SPNs, of which 549 were benign and 567 were malignant.

6 Data augmentation

To obtain the best classification accuracy of PET/CT nodule images, the initial datasets, PET/CT and LIDC-IDRI were joined. Thus, the performance of the CNNs with a variety of training sets was investigated. For the experiments, the CNNs are trained with the following sets:

- PET/CT images, referred to as Dataset_A,
- PET/CT images, augmented during the training phase (rotation by maximum of 30°, vertical – horizontal flips, width – height shift by a maximum of 4 pixels). This dataset is referred to as Dataset_B,
- Combination of PET/CT and LIDC-IDRI, referred to as Dataset_C,
- Combination of PET/CT, augmented during the training phase (rotation by maximum of 30 degrees, vertical — horizontal flips, width — height shift by a maximum of 4 pixels), and LIDC-IDRI. This dataset is referred to as Dataset_D.

6.1 CNN architectures of this study

In this section, a brief description of the benchmark of CNNs and the pretrained CNNs utilized for the classification task of this work is illustrated.

7 Benchmark of CNN architectures

The benchmark CNN architectures are designed to utilize different parameters and alternative convolution and pooling processes. The extracted features of each CNN are inserted into a Neural Network utilizing a Softmax classifier to return the probability score of each class. The optimal parameters, such as the depth of the Neural Network at the top of the CNNs, the optimizer, the receptive of the convolutional layers, the batch size and epochs of training, were defined after separate extensive separate experiments. Specifically, three experimental architectures of CNNs were developed. The developed architectures are referred to as DeepSPN (DSPN), DualDeepSPN (DDSPN), and ThreeDeepLIDC (TDLIDC). The architecture of each CNN is presented in Fig. 3.

7.1 DSPN

This CNN consists of three sets of Convolution Layers, each followed by a max-pooling layer for dimensionality reduction. The number of the filters is gradually increasing from 32, to 64, and 128. The Max Pooling layers reduce the size of the input from 32×32 to 16×16, 8×8 and 4×4, accordingly. A final Convolution Layer with 256 filters is then applied. This CNN is shallower, compared to the rest CNNs, and was developed to investigate the significance of early extracted features from the images.

7.2 DDSPN

This CNN is a dual-path network. The input image undergoes two convolution processes independently. The task of the first path is to gather local (e.g. shape, edges) information and to directly connect them to the Neural Network at the top. The second path's task is to take a more rapid step to larger filters and gather global characteristics. This CNN was developed to inspect the performance of large-scale filters, when introduced rapidly to the initial image.

7.3 TDLIDC

This CNN is similar to DDSPN, but has an extra Convolution process path applied directly to the initial input image. For the third path, a Max Pooling layer of 5×5 receptive field is applied to the input image, to reduce its size to 6×6 pixels. Then, a 512-filter Convolution Layer is applied. TDLIDC was designed to utilize both early and deep features and to investigate the rapid introduction to large filters.

8 Modified state-of-the-art Convolutional Neural networks

For the strategy of transfer learning, two state-of-the-art CNNs were utilized, which have been commonly used for medical imaging tasks.

As far as the VGG16 is concerned, this consists of 16 convolutional layers and has a very uniform architecture. The receptive field of the convolution filters is 3×3. Based on the VGG16, two CNNs were developed, referred to as VGG_{fe}, and VGG_{mod}. As far as the MobileNet is concerned, this consists of 22 depth wise separable [33] convolutions performing a single convolution on each color channel. All convolutional layers utilize 3×3, or 1×1 receptive field. For this experiment, the modification of MobileNet is referred to as $\text{MobileNet}_{\text{mod}}$. The extracted features of every CNN are supplied to a Neural Network and a Softmax classifier, to perform the classification process.

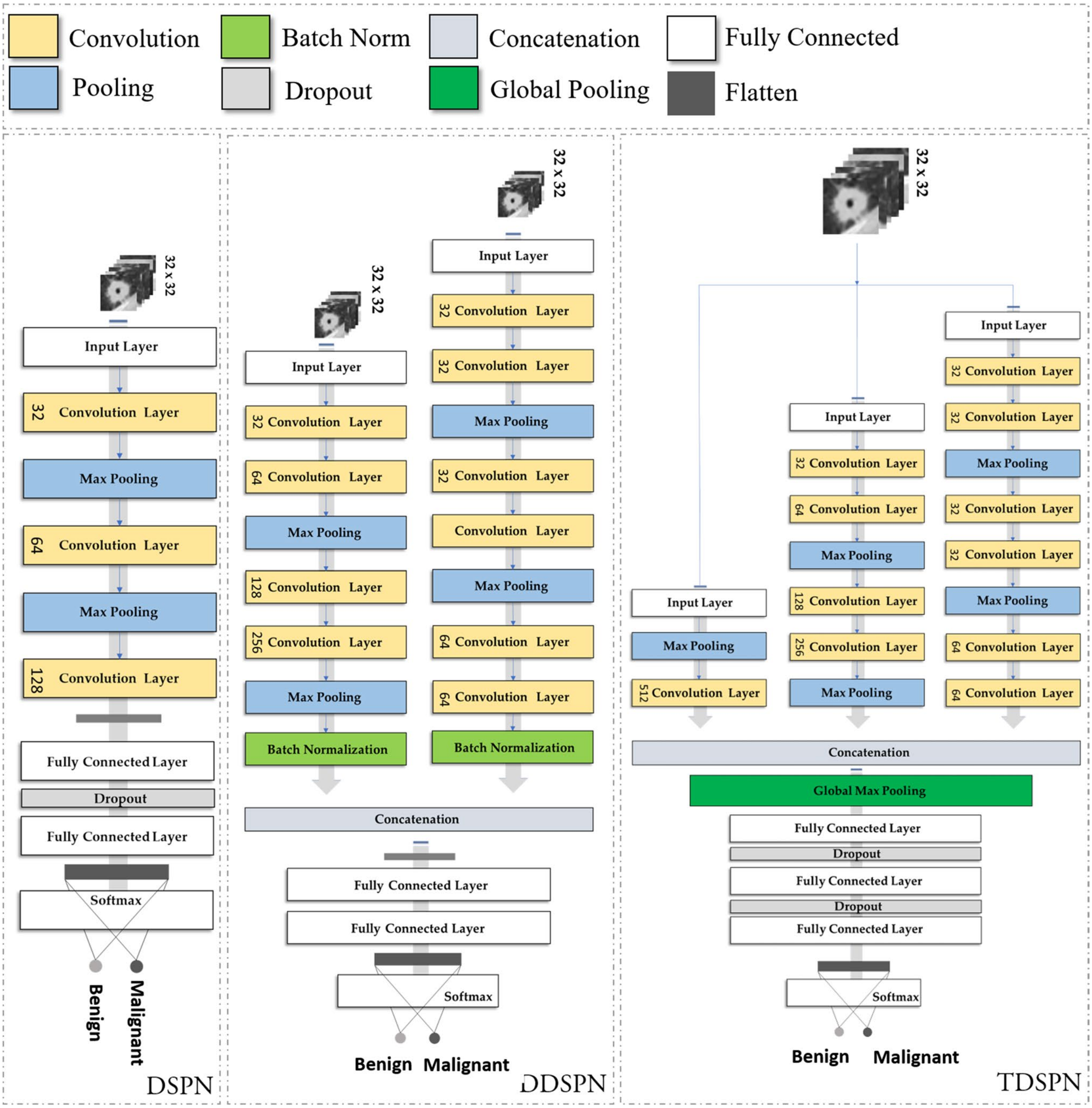

Fig. 3 Architecture and parameters of DSPN,DDSPN, and TDLIDC. The numbers inside the boxes refer to the number of filters each layer consists of

8.1 VGGfe

Concerning the VGG$_{fe,}$ the top six Convolution Layers and the two Max Pooling Layers from the VGG16 were disconnected, as they were causing a collapse of the image size, due to dimensionality reduction from the pooling layers. The rest of the layers were forced to retain their learned weights and operated as feature extractors with assigned weights.

8.2 VGG$_{mod}$

Concerning the VGG$_{mod}$, the top four layers from the VGG16 were disconnected. The first 17 layers, which extract local and possibly less significant features, were made untrainable to retain the learned weights from their initial training. The remaining layers extract more global characteristics and, thus, it was chosen to let those layers learn the global characteristics of the new images. Depending on the classification results, the strategy performance will be evaluated. Compared to VGG_{fe}, this network is allowed more freedom to adjust the weights of the convolutional layers.

8.3 $MobileNet_{mod}$

Based on MobileNet, a modification called $MobileNet_{mod}$ was constructed. The top forty-five layers were disconnected from the constructed network due to dimensionality collapse, while the remaining layers' weights were retained to perform the classic feature extraction as VGG_{fe}.

Table 1 CNNs' accuracy. The training was performed by creating four different datasets, as explained. The test set in each fold is consisted of instances only from the original PET/CT dataset ($Dataset_A$)

Network	Datasets used for training				Improvement
	$Dataset_A$	$Dataset_B$	$Dataset_C$	$Dataset_D$	
DeepSPN	0.8984	0.9001	0.8810	0.9157	+1.7%
DDSPN	0.8463	0.8637	0.8793	0.8967	+5%
TDLIDC	0.8256	0.8098	0.8194	0.8758	+5%
VGG_{fe}	0.9305	0.9175	**0.9401**	0.9184	+1%
VGG_{mod}	0.8950	0.8723	0.8819	0.8697	-2.5%
$MobileNet_{mod}$	0.8428	0.6605	0.9201	**0.9314**	+9%

9 Results

The networks were trained using $Dataset_A$, $Dataset_B$, $Dataset_C$, and $Dataset_D$, as explained. As the intention is to improve the performance and classification accuracy of the PET/CT dataset, the test accuracy is measured using folds from the original PET/CT dataset during the tenfold cross-validation procedure. The evaluation of the CNNs is based on the accuracy and AUC score, which aggregates the model's behavior for all possible thresholds to distinguish between the two classes. Table 1 presents the CNNs' accuracy. The highest values are indicated in bold.

Table 1 highlights the effectiveness of transfer learning and data augmentation methods. The highest accuracy obtained is 94% (VGG_{fe}). Every network, except VGG_{mod}, obtained its best score when trained on augmented data. The accuracy of $MobileNet_{mod}$ is increased by 9% on the dataset ($Dataset_D$) achieving the second-best performance (93%). DDSPN and TDLIDC improved their performance by 5.66 and 5%, accordingly. It is worth mentioning that VGG_{mod} was not helped by data augmentation, retaining its accuracy close to 88% in every case. The AUC scores are given in Table 2. The classification accuracy and AUC score results are also illustrated in Fig. 4 and Fig. 5

To evaluate the classification performance, some previous works are compared with the results of this work. All of these works train and test the proposed CNNs on selections of nodule images of LIDC-IDRI dataset. Since the particular study focuses on classifying a private dataset, it would be meaningful to compare the results with experiments utilizing: (a) transfer learning with pre-trained CNNs, (b) deep feature extraction approaches. To the best of our knowledge, the most notable works based on the results, the reproducibility of the experiments and utilizing transfer learning, are the first four illustrated in Table 3. The rest six mentioned approaches utilise deep feature extraction methods, but not transfer learning.

In the tables, ACC refers to accuracy, SEN refers to sensitivity, SPE refers to specificity and AUC refers to AUC score.

The proposed transfer learning scheme outperforms the rest transfer learning approaches in terms of both accuracy, sensitivity, specificity, and AUC score. Compared to the approaches, wherein only deep feature extraction without transfer learning is involved, the proposed transfer learning model obtains the highest sensitivity, which is one of the most vital metrics, since it relates the number of misclassified malignant lung nodules to the correctly classified malignant nodules.

10 Discussion

The results suggest that the optimal strategy to achieve the best classification accuracy of the PET/CT dataset involved employing VGG16 for transfer learning and expanding the training dataset by joining PET/CT and LIDC-IDRI images. Specifically, 94% classification accuracy, 92% sensitivity, 95.2% specificity, and 93.94% Area Under Curve (AUC) score was obtained. Moreover, it is verified that data augmentation, either by generating new images, or by joining the initial datasets, contributed to the performance of all the CNNs used for the experiment. Besides, the strategy of transfer learning, by utilizing a modification of the VGG16, obtains the best sensitivity, compared to several researches,

Table 2 CNNs' AUC scores

Network	Datasets used for training			
	$Dataset_A$	$Dataset_B$	$Dataset_C$	$Dataset_D$
DeepSPN	0.8974	0.8994	0.8800	0.9156
DDSPN	0.8436	0.8640	0.8809	0.8967
TDLIDC	0.8256	0.8081	0.8191	0.8729
VGG_{fe}	0.9289	0.9178	**0.9394**	0.9167
VGG_{mod}	0.8957	0.8700	0.8843	0.8691
$MobileNet_{mod}$	0.8385	0.6749	0.9185	**0.9300**

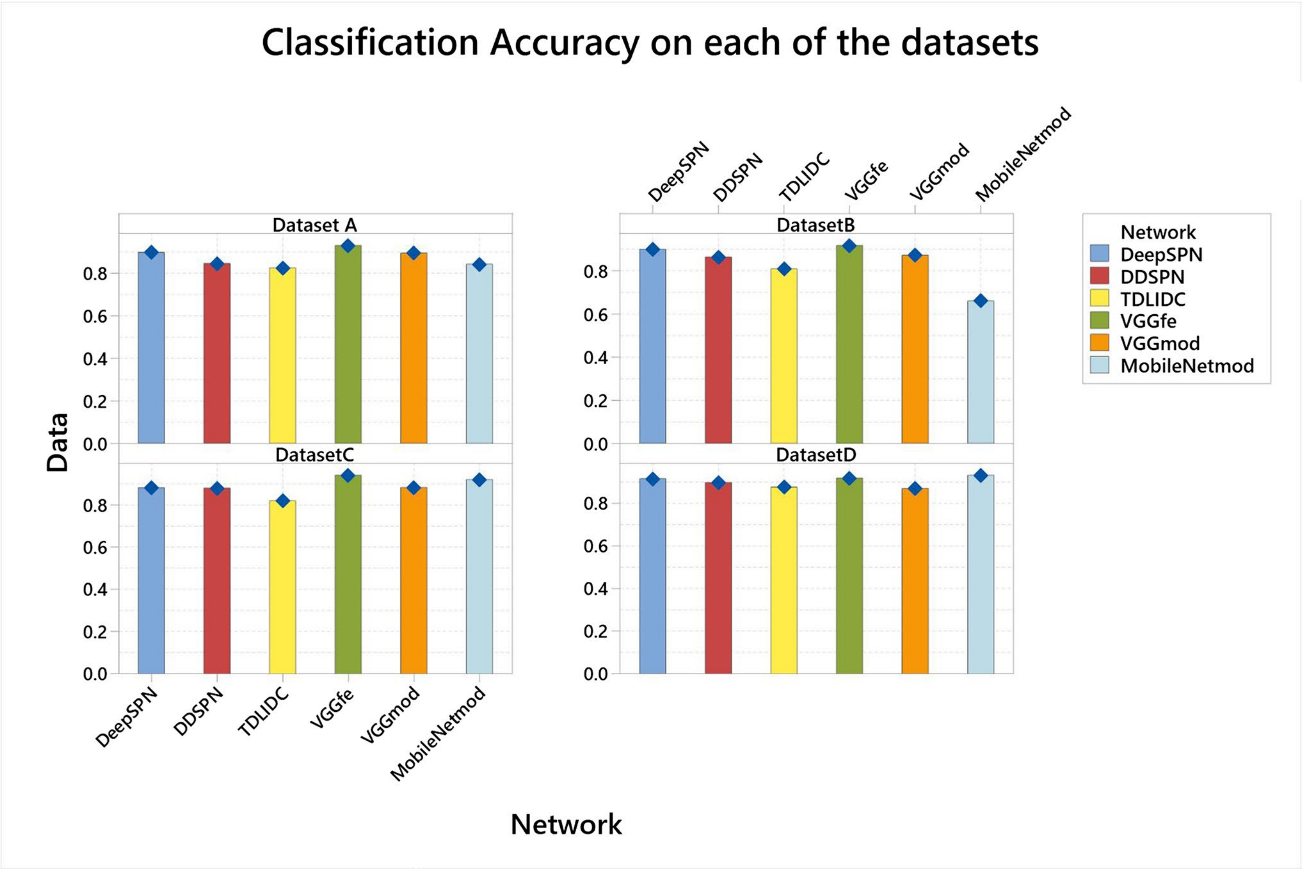

Fig. 4 Classification accuracy chart. The sub-charts correspond to the four datasets, while each bar refers to the employed CNNs

utilizing the LIDC-IDRI dataset and performing transfer learning, or training CNNs from scratch, for deep feature extraction.

Training robust and deep CNNs necessitates in the supplementation of large-scale training sets. As it is confirmed by the results, the traditional data augmentation methods (i.e. rotations, flips, shifts) aid in the expansion of the training set and slightly improve the classification accuracy. Besides, it was also demonstrated that joining CT image datasets of similar nature, further improves the abilities of the CNNs, even though image texture diversity may be present due to technical differences between CT scanners.

Transfer learning using pre-trained networks outperformed the performance of experimental CNN architectures, which were trained from scratch. Thus, it is proven that transfer learning is an effective strategy to extract representative imaging biomarkers in chest CT images.

Besides, the CNNs used as untrainable feature extractors ($\mathrm{VGG}_{\mathrm{fe}}$, MobileNet) were more effective compared to $\mathrm{VGG}_{\mathrm{mod}}$.$\mathrm{VGG}_{\mathrm{mod}}$ was allowed to learn new weights of the top layers and retain the pre-learned weights of the bottom layers. It is fair to assume that this strategy required more available data, or the data augmentation methods were not appropriate to this type of training scheme. The latter is verified by the drop of performance of VGGmod, when trained with larger datasets, containing augmented images.

Regarding the experimental CNNs and the approaches they followed to perform the feature extraction, it is clear that the rapid introduction to large filters does not perform any significant global feature collection. Specifically, the feature extraction process, followed by DDSPN and TDLIDC, was hampered by the addition of the extra convolution paths.

There are two reasons to explain this phenomenon. Either the selection of the convolutional processes of the extra paths was impeding the CNNs from gathering significant information, or the CNNs were in need of more data to complete their training. This shall be investigated in a future research.

One limitation of the current study is the restricted investigation and utilization of pre-trained networks, due to limitations of the space. There are several state-of-the-art CNNs, which were not explored. Should an even larger-scale dataset be available, more complex CNNs could be also employed, the training of which demand vaster amounts of image data.

As the available medical data increase in size rapidly, new methods for acquirement, transmission, and analysis are

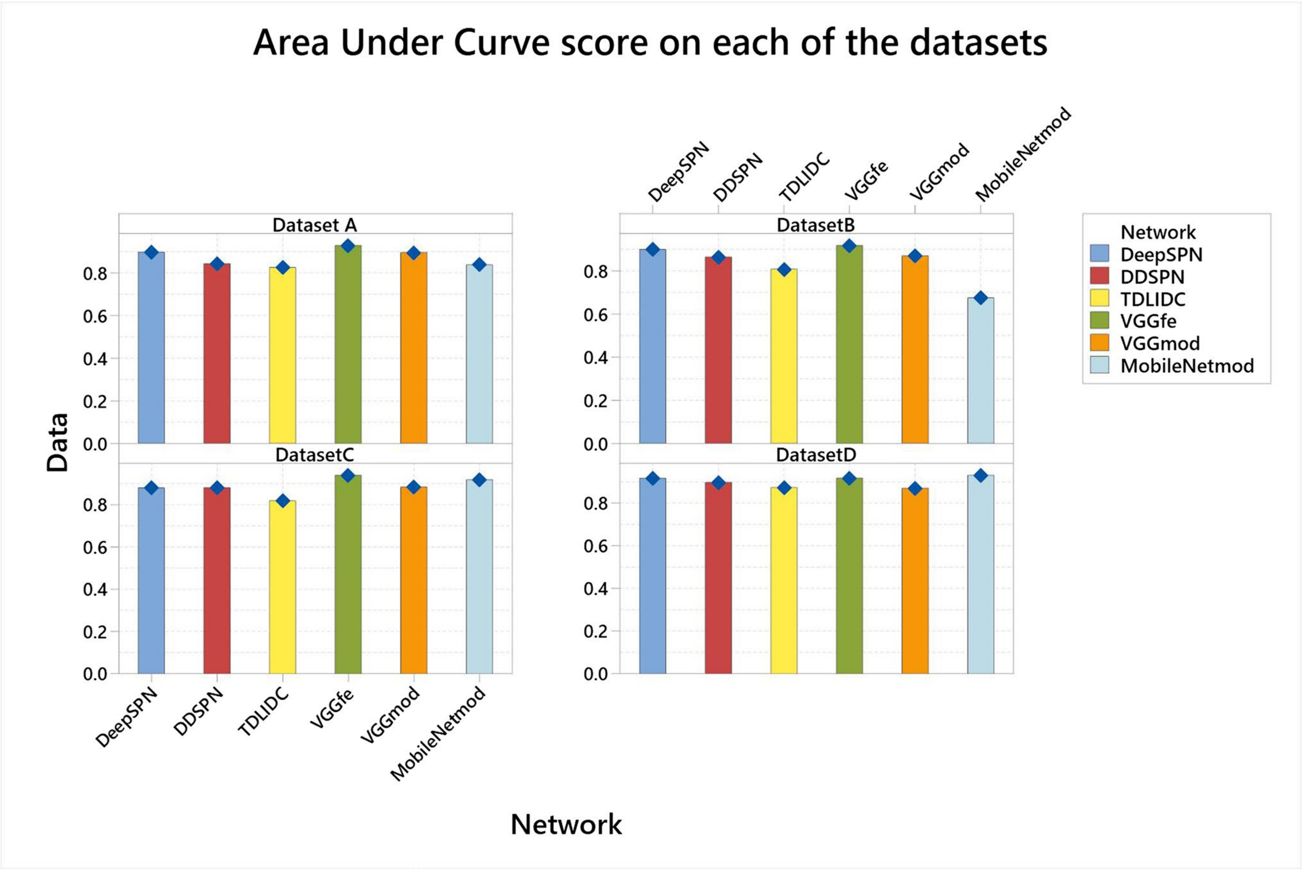

Fig. 5 Area under curve score chart. The sub-charts correspond to the four datasets, while each bar refers to the employed CNNs

becoming eminent [36–38]. More precisely, medical Image analysis involves detection, segmentation, and classification tasks [39], wherein deep learning holds and important role in the future. As it is demonstrated in the particular study, the existing state-of-the-art CNNs, although developed for non-medical reasons, can be employed for feature extraction in medical imaging.

11 Conclusion

In this paper, it was demonstrated that transfer learning is a robust and preferable strategy to circumvent the shortage of large-scale datasets to train deep and effective networks.

However, the insuffiency of data could be addressed by exploring other strategies. A future research option

Table 3 Classification performance of related researches using transfer learning. The number in parenthesis next to the test data refers to the number of nodule images used for test

Approach	Test data (size)	ACC (%)	SEN (%)	SPE (%)	AUC (%)
Zhao and Liu [15]	LIDC (743)	82.2	-	-	87.7
De Nobrega [23]	LIDC (1536)	88.4	85.4	73.5	93.2
Zhao et. al. [24]	LIDC (2028)	85	84	-	94
Xie et al. [25]	LIDC (1357)	93.4	91.4	94.1	-
Cheng [34]	LIDC (1400)	94.4	90.8	91.6	**98.4**
Wei et. al [35]	LIDC (1375)	87.1	77	93	93
Song et. al [19]	LIDC (5024)	84.1	83.9	84.3	91.6
Causey [20]	LIDC (664)	93.2	87.9	**98.5**	97.1
Dey [21]	LIDC (686)	90.4	90.5	90.3	95.5
Wu [4]	LIDC (1404)	**97.6**	-	-	-
This study	PET/CT (1152)	94	**92.7**	95.2	94

should be the investigation with more advanced data augmentation techniques, such as the utilisation of Generative Adversarial Networks [40], to generate new nodule representations in a more dynamic and sophisticated way, compared to the traditional techniques. Generating fake but realistic images, could further enhance the classification accuracy and the generalization capabilities of the CNNs.

One further direction which is of high importance, especially in areas such as medical informatics, for investigation in future research, is related to explainable – interpretable machine learning, possibly in line with E. Pintelas et al. [26, 27].

References

1. Zia ur Rehman M, Javaid M, Shah SIA, Gilani SO, Jamil M, Butt SI, (2018) An appraisal of nodules detection techniques for lung cancer in CT images. Biomed Signal Process Control 41:140–151
2. Bruno MA, Walker EA, Abujudeh HH (2015) Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 35:1668–1676
3. Wang X, Mao K, Wang L, Yang P, Lu D, He P (2019) An appraisal of lung nodules automatic classification algorithms for CT images. Sensors 19:194
4. Wu B, Zhou Z, Wang J, Wang Y (2018) Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. arXiv preprint arXiv.1802.03584
5. LeCun Y, Kavukcuoglu K, Farabet C (2010) Convolutional networks and applications in vision. In proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, France. p. 253–6
6. Zhu W, Liu C, Fan W, Xie X (2018) DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification. arXiv preprint arXiv.1801.09555
7. Russakovsky O, Deng J, et al (2015) ImageNet large scale visual recognition challenge. arXiv preprint arXiv.14090.575
8. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv.1409.1556
9. Howard AG, Zhu M, et al (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv.1704.04861
10. Zhang G, Yang Z et al (2019) An appraisal of nodule diagnosis for lung cancer in CT images. J Med Syst 43:181
11. Armato SG, McLennan G et al (2011) 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 38:915–931
12. Chen G, Zhang J et al (2019) Identification of pulmonary nodules via CT images with hierarchical fully convolutional networks. Med Biol Eng Compu 57(7):1567–1580
13. Pang S, Du A et al (2019) A novel fused convolutional neural network for biomedical image classification. Med Biol Eng Compu 57(1):107–121
14. Shin H-C, Roth HR et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35:1285–1298
15. Zhao X, Liu L et al (2018) Agile convolutional neural network for pulmonary nodule classification using CT images. Int J CARS 13:585–595
16. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359
17. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90
18. Shen W, Zhou M et al (2017) Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification. Pattern Recogn 61:663–673
19. Song Q, Zhao L et al (2017) Using deep learning for classification of lung nodules on computed tomography images. J Healthc Eng 2017:1–7
20. Causey JL, Zhang J et al (2018) Highly accurate model for prediction of lung nodule malignancy with CT scans. Sci Rep 8:9286
21. Dey R, Lu Z, Hong Y (2018) Diagnostic classification of lung nodules using 3D neural networks. arXiv preprint arXiv.1803.07192
22. Huang G, Liu Z, et al (2018) Densely Connected Convolutional Networks. arXiv preprint arXiv.1608.06993
23. Nobrega RV, Peixoto SA et al (2018) Lung Nodule Classification via Deep Transfer Learning in CT Lung Images. IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). Karlstad, Sweden, pp 244–249
24. Zhao X, Qi S et al (2019) Deep CNN models for pulmonary nodule classification: model modification, model integration, and transfer learning. XST 27:615–629
25. Xie Y, Xia Y, et al (2017) Transferable Multi-model Ensemble for Benign-Malignant Lung Nodule Classification on Chest CT. In Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S (ed) Medical Image Computing and Computer Assisted Intervention − MICCAI 2017, vol. 10435, Springer International Publishing, p. 656–64
26. Pintelas E, Livieris IE, Pintelas P (2020) A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability. Algorithms 13(1):17
27. Pintelas E, Liaskos M, Livieris IE, Kotsiantis S, Pintelas P (2020) Explainable machine learning framework for image classification problems: case study on Glioma cancer prediction. J Imaging, 6(6), 37, ID: jimaging-805312, https://www.mdpi.com/journal/jimaging/special_issues/dlmia
28. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In proceedings of the 2014 IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, p. 1717–1724
29. Kwasigroch A, Mikolajczyk A, Grochowski M (2017) Deep neural networks approach to skin lesions classification — a comparative analysis. In proceedings of the 22nd International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, Poland, p. 1069–74
30. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, (ed) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351, Springer International Publishing, p. 234–41
31. Nibali A, He Z, Wollersheim D (2017) Pulmonary nodule classification with deep residual networks. Int J CARS 12:1799–1808
32. Clark K, Vendt B et al (2013) (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26:1045–1057
33. Chollet F (2017) Xception: Deep Learning with Depthwise Separable Convolutions. arXiv preprint arXiv.1610.02357
34. Cheng JZ, Ni D et al (2016) Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Sci Rep 6:244–254

35. Wei G, Ma H et al (2018) Lung nodule classification using local kernel regression models with out-of-sample extension. Biomed Signal Process Control 40:1–9
36. Bayrakdar ME (2019) Priority based health data monitoring with IEEE 802 11af technology in wireless medical sensor networks. Med Biol Eng Comput 57(12):2757–2769
37. Michail CM, Agavanakis KN, Karpetas GE et al (2019) Information content in nuclear medicine imaging. Energy Procedia 157:1517–1524. https://doi.org/10.1016/j.egypro.2018.11.317
38. Bayrakdar ME (2019) Fuzzy logic based coordinator node selection approach in wireless medical sensor networks. In 2019 4th International Conference on Computer Science and Engineering (UBMK) (pp. 340–343). Presented at the 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey: IEEE.
39. Mohammed Z, Abdulla A (2020) Thresholding-based white blood cells segmentation from microscopic blood images. UHD J Sci Technol 4(1):9
40. Goodfellow I, Pouget-Abadie J, et al (2014) Generative adversarial nets. In Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, (ed) Advances in neural information processing systems 27, Curran Associates, p. 2672–2680

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ioannis Apostolopoulos received his Bachelor and Master Degree in 2017 from the Department of Electrical and Computer Engineering, University of Patras. He is a Ph.D. student at the Department of Medical Physics, School of Medicine, University of Patras. His research interests include Deep Learning and Biomedical Engineering.

Emmanuel Pintelas received his Bachelor Degree in 2018 from the Department of Electrical & Computer Engineering, University of Patras. Currently, he is a Ph.D. student in Department of Mathematics, University of Patras. His research interests include machine learning, deep learning and their applications.

Dr. Ioannis E. Livieris received his B.Sc., M.Sc. and Ph.D. Degrees in Mathematics from the University of Patras, Greece in 2006, 2008, and 2012, respectively. He is currently an Adjunct Professor at the University of Peloponnese. His research interests include neural networks, data mining, machine learning, and their applications.

Professor Dimitris I. Apostolopoulos has a PhD in Nuclear Medicine from the University of Athens. His interests are mainly focused on the clinical applications of Hybrid Imaging (PET/CT and SPECT/CT).

Nikolaos Papathanasiou is an Assistant Professor of Nuclear Medicine/Hybrid Imaging working in the PET/CT Unit of the Patras University Hospital. He is a graduate of the University of Athens, Medical School and holds an MSc in Biostatistics. His research interests include multimodality oncologic imaging in lung cancer, neuroblastoma and neuroendocrine tumors and imaging in movement disorders.

Dr. Panagiotis Pintelas Professor of Computer Science in the Department of Mathematics at the University of Patras, Greece. His research interests include Software Engineering, AI and ICT in Education, Machine Learning and Data Mining. He was involved in or directed several National and European Research and Development projects.

George Panayiotakis is a Professor of Medical Physics and Director of the Department of Medical Physics, School of Medicine, University of Patras, Greece. His research interests are focused in Medical Radiation physics and specifically medical imaging, medical image quality, quality assurance, dosimetry, radiation protection, medical radiation detectors and simulation of medical imaging systems. He has published over 230 articles in International peer-reviewed journals. He has co-authored 2 books and 8 book chapters. He has held various positions including that as Rector of the University of Patras (2010-2014).