Benign-malignant pulmonary nodule classification in low-dose CT with convolutional features

Mehdi Astaraki^{a,b,*}, Yousuf Zakko^c, Iuliana Toma Dasu^{b,d}, Örjan Smedby^a, Chunliang Wang^a

^a KTH Royal Institute of Technology, Department of Biomedical Engineering and Health Systems, SE-14157 Huddinge, Sweden

^b Karolinska Institutet, Department of Oncology-Pathology, Karolinska Universitetssjukhuset, Solna, SE-17176 Stockholm, Sweden

^c Karolinska University Hospital, Imaging and Function, Radiology Department, Solna, SE-17176 Stockholm, Sweden

^d Stockholm University, Department of Physics, SE-106 91 Stockholm, Sweden

ARTICLE INFO

Keywords:

Pulmonary nodule

Benign-malignant classification

Deep features

ABSTRACT

Purpose: Low-Dose Computed Tomography (LDCT) is the most common imaging modality for lung cancer diagnosis. The presence of nodules in the scans does not necessarily portend lung cancer, as there is an intricate relationship between nodule characteristics and lung cancer. Therefore, benign-malignant pulmonary nodule classification at early detection is a crucial step to improve diagnosis and prolong patient survival. The aim of this study is to propose a method for predicting nodule malignancy based on deep abstract features.

Methods: To efficiently capture both intra-nodule heterogeneities and contextual information of the pulmonary nodules, a dual pathway model was developed to integrate the intra-nodule characteristics with contextual attributes. The proposed approach was implemented with both supervised and unsupervised learning schemes. A random forest model was added as a second component on top of the networks to generate the classification results. The discrimination power of the model was evaluated by calculating the Area Under the Receiver Operating Characteristic Curve (AUROC) metric.

Results: Experiments on 1297 manually segmented nodules show that the integration of context and target supervised deep features have a great potential for accurate prediction, resulting in a discrimination power of 0.936 in terms of AUROC, which outperformed the classification performance of the Kaggle 2017 challenge winner.

Conclusion: Empirical results demonstrate that integrating nodule target and context images into a unified network improves the discrimination power, outperforming the conventional single pathway convolutional neural networks.

Introduction

Lung cancer is the leading cause of cancer-related death worldwide [1]. Although the 5-year survival rate of locally advanced lung cancer is less than 5%, fortunately, early diagnosis of lung cancer when the tumor is small and asymptomatic can significantly improve the 5-year survival rate to more than 60% [2]. This has encouraged a large number of lung cancer screening trials and most have demonstrated benefit in screening persons at high risk. Low-Dose Computed Tomography (LDCT) can capture high-resolution details of the lungs and surrounding tissues and has been recognized as a standard imaging modality for lung cancer screening. In an American National Lung Screening Trial, 39.1% of participants were diagnosed with at least one pulmonary nodule [3]. The term “pulmonary nodule” refers to a moderately well-margined

round opacity with the largest diameter less than 3 cm and can be benign or malignant [4]. Most lung cancers emerge from small malignant nodules. However, only approximately 20% of all subjects with nodules represent lung cancer [5]. There are various subtypes of lung nodules, such as solid, part-solid, and ground glass nodules with different cancer probabilities [6,7]. To assess the likelihood of nodule malignancy, radiologists examine the CT images on a slice-by-slice basis and follow the supporting guidelines such as LungRADs [8] and Fleischner [9], which rely only on the morphological characteristics of the nodules, e.g., size [6,9,10]. Derived malignancy likelihood from imaging scans may be used to triage patients for further investigations such as follow-up scans with LDCT, PET-CT, or even sampling biopsy [8,11]. Besides nodule size, it has been verified that the nodule’s intensity distribution, as well as its relative position, are strongly related to

* Corresponding author at: Hälsovägen 11C, SE-14157 Huddinge, Sweden.

E-mail address: mehast@kth.se (M. Astaraki).

<https://doi.org/10.1016/j.ejmp.2021.03.013>

Received 27 November 2020; Received in revised form 1 March 2021; Accepted 5 March 2021

Available online 25 March 2021

1120-1797/© 2021 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

lesion malignancy [5]. The complex multivariant diagnosis criteria and similar visual attributes shared between some cases of benign and malignant nodules make it challenging even for experienced radiologists to discriminate them (Fig. 1B). Developing a computer-based assistant tool for radiologists to objectively classify the characteristics of lung nodules to support clinical decision on future interventions remains a critical need.

Conventional Computer-Aided Diagnosis (CAD) systems focus on extracting hand-crafted features (i.e. defined by the inventor of the system) from lung nodules to train machine learning models for automatic nodule classifications. In this context, distinct feature sets such as Histogram of Oriented Gradients [12], Local Binary Patterns (LBP) [13], and radiomics [14–18] were extracted to train learning algorithms like Support Vector Machines [19], Linear Discriminant Analysis [20], and Random Forest (RF) [21]. Although these schemes lead to promising results when dealing with well-defined nodules, a major disadvantage of such approaches is their limited ability when the nodules appear with various shapes, sizes, and contexts [22]. Convolutional Neural Networks (CNNs) have emerged as a powerful alternative solution for nodule classification tasks [23]. They use an end-to-end training scheme, i.e., the entire image or image patch is fed into a network while getting a classification label as output. CNNs automatically learn to extract useful image features by adjusting the weights of their convolution kernels and therefore eliminate the need for human-dictated feature engineering. The fact that CNNs adaptively learn the optimal representations in an entirely data-driven scheme [24] by capturing the spatial dependency in images through the application of relevant features helps them to outperform classical CAD systems.

In recent years, numerous deep learning-based models have been developed to address the problem of classifying the malignancy of lung nodules. In [25] the authors fine-tuned a pre-trained 3D CNN and then fused the complementary information from six high-level nodule attributes via a multi-task learning framework. In DeepLung [26], a gradient boosting machine with 3D dual-path network features was proposed to

classify the detected nodules. Although the 3D CNNs yielded promising results, learning a large number of hyperparameters associated with 3D networks from a limited number of medical images remains a challenge. As an alternative, accurate nodule classification results have also been achieved using patch-based multi-view slices to train 2.5D CNNs [5,27–29]. Along with powerful end-to-end classification networks, another family of approaches utilizes deep features extracted from unsupervised Auto-Encoder (AE)-like reconstruction networks followed by a supervised learning algorithm [30–36]. It is also worth mentioning that high malignancy classification accuracy has been reported by combining quantitative hand-crafted features with CNN-based features [37,38].

Despite the abundance of publications aiming at nodule classification in LDCT, no efforts, to the best of our knowledge, have been put into assessing the nodule target and nodule context information separately. Moreover, a large number of previous studies have investigated the performance of their proposed models on the publicly available LIDC-IDRI dataset [39]. As accurate classification results have already been achieved on that dataset [29,38], it is necessary to provide another challenging large-scale dataset for external validation.

In this paper, in order to design a robust model for pulmonary nodule malignancy prediction, we propose to train dual pathway 3D-CNN models fed with nodule target images (region inside) and nodule context images (region inside and around) simultaneously. By masking out information either inside the nodule or outside the nodule, we expect to force the convolutional pathways to learn different image features for different regions and disentangle the two aspects of nodule characteristics explicitly. By instructing CNNs to learn and extract image features from inside and outside the nodule images separately, we aim to efficiently capture both intra-nodule heterogeneities and contextual information with CNNs.

Fig. 1. 2D axial view examples of pulmonary nodule appearance in LDCT images. A) All the scans were manually examined by an expert radiologist slice to detect the nodules (yellow bounding box), and the nodule boundaries were delineated (red contour) in the cropped patches. B) Illustration of similar visual characteristics between benign (first row) and malignant (second row) nodules. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Material and methods

The proposed model consists of two major pathways to process the paired of nodule target and context images simultaneously. These two pathways were implemented with two alternative options, representing two different strategies for deep feature extraction: supervised models including VGG-like [40], ResNet [41], DenseNet [42], EfficientNet [43], and a Variational Autoencoder (VAE) [44] as an unsupervised model. Each network is constructed on the basis of two similar parallel sub-networks with a shared layer to extract features from the paired images. The extracted deep features (the outputs of the second last layers of a CNN) are then augmented to tackle the challenge of working with an imbalanced data set by employing the Synthetic Minority Oversampling Technique (SMOTE) [45] before training a learning algorithm. A schematic illustration of the study steps is shown in Fig. 2. An RF algorithm is employed on top of the trained models to classify the extracted feature vectors as either benign or malignant. In addition, conventional single-pathway models with similar architectures as one of the parallel networks in dual-pathway models were used to train nodule context and nodule target images separately, followed by separate feature augmentations and different RF classifiers.

Patient data

The Kaggle Data Science Bowl 2017 challenge aimed at automatic solutions for lung cancer diagnosis in LDCT images [46]. The challenge dataset comprised a total number of 2101 axial LDCT images from high-risk patients in DICOM format. This data includes 1397, 198, and 506 subjects for training, validation, and testing, respectively. Each image was labeled as either “1”, if the patient was diagnosed with lung cancer within one year from the scan, or “0” otherwise. This dataset was collected from many different centers that consist of large variations in image quality and acquisition parameters. More specifically, the number

of axial slices among subjects varies between 94 and 541, slice thickness varies in the range of 0.625 to 2.50 mm, and voxel resolution, along X and Y axes, varies from 0.490 to 0.957 mm. It should be noted that the coordinates/segmentation masks of the nodules were not provided.

In this study, the training set of the challenge was analyzed. In particular, of 1397 training subjects, 968 LDCTs were examined by an expert radiologist, and 1297 pulmonary nodules were identified and segmented manually by using MiaLab software [47]. By using the clinical follow-up labels, 876 nodules were marked as benign, and the rest of 421 segmented nodules were denoted as malignant ones. Besides, an image patch that best covers each nodule was cropped and extracted. Accordingly, in this paper, the term “nodule target image” indicates the segmented nodule image, and “nodule context image” refers to a patch that covers the segmented nodule as well as its nearby regions.

Image preprocessing

Prior to feature extraction, all the cropped patches were pre-processed in three steps. Considering the fact that the voxel size among the employed dataset varied significantly and the smallest one was $0.49 \times 0.49 \times 0.625 \text{ mm}^3$, first, original patches were resampled isotropically to a unified inner plane spacing as 0.2 mm^3 using a bicubic interpolation function. Then the intensity ranges were clamped to $[-1000, 500]$ Hounsfield Unit. A further step was applied in which patches were rescaled by zero-padding the original sizes into $128 \times 128 \times 128$ voxels followed by intensity normalization in the range of $[0, 1]$.

Deep feature extraction

Besides training CNNs end-to-end to make the classification directly, in this paper, we developed a model that consists of two major components: a deep feature extractor and a classification component. In

Fig. 2. Schematic illustration of the study pipeline. Nodule context and target images were used, separately, to train single pathway VGG-like networks. The same network architectures as single-pathway networks were employed to design a dual-pathway network that is fed by both the context nodule and target nodule images. Each model was trained with an end-to-end learning scheme. From the trained models, deep features were extracted and augmented using the SMOTE algorithm to train random forest classifiers. This pipeline is applied to both the supervised and unsupervised learning schemes detailed in the *deep feature extraction* section.

particular, the output of the second last layer (the layer before the final classification labels) of a CNN model is extracted as deep features. Then, a learning algorithm as the second component is added on top of the extracted deep features to generate the classification results.

Supervised deep feature extraction

In the supervised-based nodule classification setup, the proposed dual pathway model was applied to different model architectures. The two input branches are fed with the nodule context and the nodule target images separately. In other words, while the nodule target pathway is assumed to mainly learn the association between the intra-nodule representations and class labels, the role of the nodule context pathway is to primarily learn the correlations between the context information and class labels. Thereafter, by concatenating the learned features from each of the pathways in a last shared layer, the model will learn to predict the class labels by adaptively learning the disentangled intra- and context-nodules attributes simultaneously. This structure was applied to the employed VGG-like, ResNet, DenseNet, and EfficientNet models.

The unified dual-pathway VGG-like model [40] consists of two convolutional pathways for representation learning, each followed by two dense layers and a shared final dense layer (Fig. 2). The convolutional backbones of the network architectures are the same and comprise five 3D convolutional blocks, each of which contains two convolutional layers, two batch normalization layers, a max-pooling, and a dropout layer as well; except for the last block that does not include max-pooling layer. The base block starts with eight convolutional filters, and this number increments by a factor of two as the blocks go deeper into the model. Two dense layers were connected to the last convolutional blocks with 2048 and 1024 neurons, respectively. Accordingly, the final shared layer contains 2048 neurons.

Each pathway of the ResNet model [41] starts with a 3D convolutional filter containing four feature maps, followed by four convolutional blocks that consist of 6, 8, 12, and 6 residual modules and dropout layers, respectively, and ends with a global average pooling and a dense layer. Two max-pooling operators were added after the first convolutional layer and the third convolutional block. The outputs of the last residual module of each pathway were concatenated before feeding the global average pooling operator to form the shared layer.

Each DenseNet [42] pathway filters the images with a 3D convolutional filter with eight feature maps followed by four dense blocks and four transition layers. Each dense block consists of eight layers; the growth rate of the feature maps was set as 10, and the compress factor of 0.5. The transition layers used in these experiments consist of a batch normalization layer, $1 \times 1 \times 1$ convolutional layer followed by a drop out layer, and a $2 \times 2 \times 2$ average pooling layer. The output of the convolution filter of the last transition layers of the parallel networks was concatenated to form the shared layer between the context and target paths.

Instead of scaling up only one network attribute out of resolution, width, and depth, the EfficientNet [43] model was proposed to efficiently scale them up all together with a strategy known as compound scaling. Each pathway of the employed model begins with a 3D convolutional filter, followed by five Inverted Residual Blocks (IRB) with 16, 24, 40, 80, and 112 feature channels, respectively. ReLu activations were replaced by SeLu activation functions, and squeeze and excitation modules were added within each block to weigh each feature channel instead of treating them equally. A global average pooling operator was added at the top of the network before the final dense layer. The output of the last IRB of the parallel networks was concatenated to form the final shared layer in the dual-pathway model.

Unsupervised feature extraction

Unlike end-to-end supervised classification networks, VAE is a scalable generative model that is able to capture the rich distribution of high dimensional data via backpropagation [48]. Let x be the input image,

and z indicates the latent code of a VAE with a deep CNN-based encoder and decoder networks. If $p(z)$ represents the imposed prior distribution on the codes, $q(z|x)$ and $p(x|z)$ will be the encoding and decoding distributions, respectively. Also, let $p_d(x)$ be the distribution of input data and $p(x)$ the distribution of the model. Then, the encoding function of the AE, $q(z|x)$, implies an aggregated posterior distribution of $q(z)$ on the latent codes of the AE as follows [31]:

$$q(z) = \int_x q(z|x)p_d(x)dx$$

VAEs are powerful models to find out how latent variables affect output. The extracted features from the encoder part of the VAE provide high-level semantic attributes not only capture the different conventional semantic nodule features like lobulation, spiculation, etc., into account but are also able to capture the association between them [32,35].

The architecture of the encoder part of the VAE consists of four 3D convolutional blocks, each of which contains two convolutional layers followed by batch normalization, as well as a max-pooling layer. A dense layer with 1024 neurons was connected to the last convolutional block and worked as a deterministic feature extractor. The model is completed by a sampling layer, followed by a symmetric decoder with four transpose convolutional blocks. The same architecture was employed for nodule targets and nodule context images separately.

Model training

All models were analyzed with a total number of 1297 patches with a size of $128 \times 128 \times 128$ in a 5-fold cross-validation fashion. 3D image augmentation techniques were applied during the training of deep networks to minimize the risk of overfitting. In specific, the augmentation methods include: randomly flipping the nodules along with one of the three-axis and affine transformation as well. Both supervised and unsupervised models were trained for 150 epochs during each fold of cross-validations with Adam optimizer and an initial learning rate of 0.0001. The conventional binary cross-entropy loss function was employed for training the supervised models, and per-voxel l_2 loss regularized by Kullback-Leibler divergence organized the loss function of the VAE model.

Nodule classification

As the final step, the extracted deep features are employed to perform the nodule malignancy prediction. In particular, 1024 deep features extracted from the nodule target images combined with 1024 deep features from nodule context patches to construct the feature pools.

Experimentally, RF was chosen as the classifier. However, since the data set is highly imbalanced with the malignant group as the minority class, the SMOTE algorithm is employed to augment the minority class by synthetically generate new data. In other words, it selects k -(variable) closing neighbors in the feature space from the minor class, draws a line (hypercube) between them, and generates a new sample along that line (hypercube). As a result of this step, 455 new feature vectors belonging to the malignant class were synthesized and increased the total number of data to 1752 to train an RF classifier in a 5-fold cross-validation approach. The performance of the analyses was assessed using the area under the receiver operating characteristic curve (AUROC) over the validation sets which were not employed in the model training. In order to statistically test the significant difference between the computed AUROC values and random performance (AUROC = 0.5), a permutation test was employed in which the class labels of the corresponding deep feature vectors were randomly permuted before training RF classifiers. Furthermore, to test the statistical significance in a pairwise comparison of target, context, and combined nodule images, the method proposed by Delong [49] was utilized.

Additionally, in order to specify the advantage of the proposed method, comparisons were conducted by training the similar models in

an end-to-end fashion. In other words, the same single-pathway architectures of the proposed supervised models were used to train, separately, nodule target and nodule context images but without using SMOTE as feature augmentor and RF classifiers. Likewise, end-to-end dual-pathway supervised models were employed to train nodule context and target images simultaneously. In the rest of the paper, the term “baseline” refers to such models.

In order to quantify the effect of the feature augmentation step, the extracted deep features were employed to train the RF learning model but without using the SMOTE algorithm. Furthermore, to investigate how realistic the synthesized features generated by SMOTE method, the augmented features were only used to train the RF models, and 20% of the real extracted features were employed as testing sets. Finally, a set of experiments with an end-to-end baseline model as well as an augmented feature model were conducted by analyzing only the regions around the nodules, i.e., the nodules were masked out from the context images. All the models were implemented with Scikit-learn [50] and TensorFlow [51] libraries.

Results

The discrimination power of the baseline models, deep supervised and unsupervised features extracted from nodule target and context images, and their integrations are reported in the following sections.

The prediction power of the end-to-end trained baseline models, separately, for nodule target and context images, as well as their combinations (dual-pathway), is presented in Table 1. Table 2 represents the classification power of the augmented extracted deep features (See Fig. 1 in supplementary material). It can be observed that for both the nodule context and target images, as well as their combinations, the prediction powers of end-to-end baseline models are inferior to the augmented features. From both tables, it can be seen that the AUROC values of context nodules are slightly higher than that of the target nodules, except for two cases of VGG and ResNet baseline models. Among the supervised models, either as the baseline or augmented features, DenseNet performs relatively better than the others.

In general, we observe that both the deep feature approach (supervised and unsupervised) outperforms the end-to-end baseline models. In the meantime, the supervised dual pathway setup seems to outperform the unsupervised dual pathway with a rather distinct margin. Table 2 (first row) shows the results of malignancy prediction with semantic features extracted from nodule context and target images as well as their combination via the unsupervised VAE model. With this model, it can be seen that nodule context images carry more prediction power than the nodule target images; however, their combination is not constructive and does not guide any improvements. Similar to the unsupervised model, supervised context features are more informative than target attributes. Moreover, interesting results achieved from the supervised models where the combination of context and target features resulted in a slightly higher AUROC value than each of them separately, which

Table 1

The prediction power of the end-to-end supervised baseline models. The highest value on each row is marked in bold. Lower and upper limits of confidence interval at 95% level are indicated in square brackets. All AUROC values were significantly different from 0.5 (randomness).

Baseline Model	Prediction Performance (AUROC)		
	Target	Context	Dual-pathway
VGG	0.801 [0.777,0.824]	0.795 [0.774,0.816]	0.821 [0.794,0.831]
ResNet	0.785 [0.756,0.806]	0.763 [0.740,0.782]	0.794 [0.771,0.815]
DenseNet	0.792 [0.775,0.813]	0.806 [0.788,0.827]	0.824 [0.798,0.837]
EfficientNet	0.783 [0.759, 0.809]	0.798 [0.772,0.818]	0.808 [0.784,0.834]

Table 2

The prediction power of the proposed augmented deep features. The highest value on each row is marked in bold. Lower and upper limits of confidence interval at 95% level are indicated in square brackets. All AUROC values were significantly different from 0.5 (randomness).

Feature Type	Prediction Performance (AUROC)		
	Target	Context	Dual-pathway
VAE	0.851 [0.837,0.866]	0.868 [0.851, 0.883]	0.855 [0.839,0.871]
VGG	0.898 [0.882,0.913]	0.917 [0.901,0.928]	0.920 [0.905,0.934]
ResNet	0.902 [0.886,0.917]	0.903 [0.886,0.918]	0.909 [0.895,0.923]
DenseNet	0.906 [0.890,0.921]	0.924 [0.908,0.938]	0.936 [0.921,0.950]
EfficientNet	0.905 [0.890,0.919]	0.927 [0.912,0.940]	0.931 [0.917, 0.944]

points to the complementary role of target and context features. Tables 3 and 4 in supplementary material indicate the pairwise statistical comparison between the calculated AUROC values for nodule target and context images as well as their combinations.

Feature fractioning

To study the effect of the size of feature pools on the RF learning algorithm, experiments were conducted by exposing different fractions of the feature sets. Specifically, the experiments performed on 25%, 50%, and 70% of all the extracted features as the training feature sets, and the prediction power is reported over the remaining features as test sets (Table 3). As was expected, increasing the size of the training feature sets improved the prediction power. The results show that even with a smaller size of the training feature set, supervised CNN-based features outperformed unsupervised features.

Feature augmentation effect

Table 1 in supplementary material shows the quantified effects of the feature augmentation step. The imbalanced set column of this table represents the results of the raw extracted VGG features, without being augmented, that were employed to train an RF model. Comparing against the baseline model (VGG row in Table 1), one can behold that the end-to-end trained model performs better than extracting the features and then, train another learning algorithm. Furthermore, comparing the results of the balanced set of Table 1 in supplementary material against the VGG row in Table 2, we can observe that employing both the synthesized and real extracted features in the training set increases the model performance (AUROC = 0.873 vs. AUROC = 0.920).

Finally, the results of masking out the nodule information from the images and only analyzing the regions around the nodules are presented in Table 2 in the supplementary material. In general, these images carry less discrimination power than both target nodule images and context nodule images.

Table 3

The effect of the size of the training feature set on the prediction power (AUROC). The reported results were calculated over the test feature sets after 5-fold cross-validation. For each fraction, the highest value is marked in bold. Lower and upper limits of confidence interval at 95% level are indicated in square brackets. All AUROC values were significantly different from 0.5 (randomness).

Fraction of Training Features	Feature Set	
	Unsupervised (VAE)	Supervised (VGG)
0.25	0.795 [0.779,0.808]	0.869 [0.852,0.883]
0.50	0.817 [0.801,0.831]	0.881 [0.866,0.895]
0.70	0.836 [0.823,0.850]	0.894 [0.880,0.909]

Discussion

One of the most demanding tasks that radiologists need to carry out is to identify malignant pulmonary nodules from benign ones when reading chest LDCT images. This manual step is cumbersome, subjective, time-consuming, and is often associated with considerable inter/intra-observer variability. Numerous computer-assisted tools have been introduced based on image processing, conventional machine learning, as well as advanced CNN models. CNN-based features are indeed abstract, and there is no guarantee that the black-box models with limited supervision derive the classification results by looking at the nodule lesions [38,52]. Moreover, previous studies have shown that separating image features extracted from inside the nodule and from the region immediately outside the nodule could lead to better classification accuracy. More specifically, the following findings have been claimed: (1) intra-tumor heterogeneity that implies a strong correlation with malignancy can be captured with textural features [18,53]; (2) contextual characteristics of the region where the nodules interact with nearby tissues represent important characteristics of the lesions [54].

Our hypothesis was that the same practice could also help CNNs to better distinguish malignant from benign lung nodules. The main objective of this study has been to investigate this hypothesis to find out whether instructing CNNs to learn and extract image features from inside and outside the nodule separately may allow the networks to learn disentangled characteristic features, and thus lead to better classification accuracy. In order to allow the CNNs to effectively distinguish intra-nodule characteristics from contextual nodule attributes, deep representation learning was achieved by training dual pathway models fed by nodule context and target images separately. To conduct a more thorough evaluation, the proposed dual pathway models were trained both with supervised and unsupervised learning approaches, and different combinations of the extracted features were used for malignancy prediction. In total, three separate pipelines were developed and investigated to predict the lung nodule malignancy from LDCT images. The findings of our study suggest that deep representative features have the potential to capture different types of nodule attributes. Our proposed approach for integrating nodule target and context images via dual pathway networks is, to the best of our knowledge, the first attempt to develop such integrated lung nodule malignancy prediction models.

The design of dual pathway deep architectures is based on the idea that both the intra-nodule heterogeneity and context information have a strong correlation with nodule malignancy. In particular, visual characteristics of the nodules, such as heterogeneity within intensity and textures, spiculation, and presence of cavities can be quantified from the nodule target images. In contrast, other important factors, such as attaching to the artery, pleura, and chest wall, can only be captured from context images. Instead of using complicated multi-scale patches [55] or multi-view CNNs [27,28], which are sensitive to shape characteristics, we utilized the 3D patches centered at nodules to train the 3D networks and applied the data augmentation techniques to mitigate the tendency of the network to learn entangled images features that increase the risks of overfitting on the training data. Our experiment results seem to support this hypothesis, with the dual pathway setups slightly improving the classification accuracy. Comparing the classification performance achieved in this study against the Kaggle Data Science Bowl 2017 challenge winner [56], one can see that while the performance of the baseline end-to-end models are inferior to that of the challenge winner, the proposed method of balancing the minority class regardless of the complexity of the model backbones outperformed the best performance of the challenge with 3 percent of improvement ($AUC_{Dual-DenseNet} = 0.93$ vs. $AUC_{winner} = 0.90$). It should be mentioned that in [56] they, first, detected the abnormalities via a 3D CNN, then employed the highest confidence of detections along with manual labelling to classify the malignancy, while nodule detection is not included in our evaluation setup.

Another advantage of CNNs is that they do not necessarily need the

nodule target images. It can thus be seen that context images contain more prediction power than target images. This could be explained by the fact that context images represent both the relative location of the nodule with respect to nearby structures and intra-nodule attributes. In contrast, masking out the nodules from images and analyze only the regions around the nodules loses the important intra-nodule characteristics that lead to inferior performance compared against context and target nodule images. Furthermore, experiments show that supervised CNN-based features have higher prediction power than unsupervised ones. In fact, training a supervised network is assumed to capture more relevant features with respect to the classification task than an unsupervised network, which is likely to learn the latent semantic representations to reconstruct a blurry version of the input images. In other words, training a supervised classifier drives the model to learn more informative representations concerning the class labels rather than an unsupervised reconstruction network which is trained to capture the distribution of latent semantic features. Although the performance of the unsupervised model was inferior compared to supervised models, one potential application of such models would be to train dataset with only a limited number of labeled subjects to extract deep features that, later, can be used to train non-data-greedy classification models.

Previously tested approaches to reducing the effect of class imbalance in network training include, but are not limited to, adding class weights to the loss function and oversampling from the minority class (malignant) per batches. Employing SMOTE as an augmentation technique to generate synthetic samples in the feature space helped to further improve the classification performance by preventing the training bias of the RF classifier toward the majority (benign) class. The importance of such a feature augmentation step is highlighted by the almost 11 percent AUROC improvement of the deep feature model ($AUROC_{DenseNet-Augment} = 0.936$) against the end-to-end trained baseline model ($AUROC_{DenseNet-Baseline} = 0.824$). Furthermore, the declined performance of the learning algorithm after using the augmented features only for training the RF model implies the fact the new synthesized features were not just replications of the original features; therefore, this process does not lead to the problem of the model overfitting. Additionally, the impact of fractioning the size of the training set on deep features was not remarkably dramatic. In particular, reducing the size of supervised deep features to 25% of the entire feature set diminished the model performance from 0.920 to 0.869.

Despite the encouraging results achieved in this study, there are a few noticeable limitations. First, in this study, manually segmented pulmonary nodules by only one radiologist expert were analyzed. Having more than one segmentation mask, however, would reduce the risk of inter-observer variability. Moreover, the full lung cancer screening pipeline would benefit from automatic detection of the nodules as well. This can be addressed in future studies by developing a nodule detection model followed by the classification pipeline. Second, the generalization of the proposed approach could be further strengthened if the learned models were employed as pre-trained networks from other external datasets such as LIDC-IDRI. Last but not least, the manually extracted visual radiological features can also be incorporated along with quantitative radiomics as well as deep abstract features.

Conclusion

In this paper, we proposed a dual pathway deep neural network to learn disentangled image features to classify benign and malignant lung nodules. This is achieved by feeding the nodule target and context images to different convolutional pathways. Our empirical results show that the dual pathway architecture performs slightly better than the conventional single pathway CNNs, and the supervised CNN-based features carry more discriminatory power than the unsupervised extracted features. Our experiments verified that the integration of nodule context with target images could successfully capture and discern the intricate characteristics of the nodules. Future work should

include employing longitudinal scans to quantify the changes in nodule characteristics for a more accurate prediction of malignancy.

Data availability

The dataset used and analyzed in this study is available from the corresponding author on a reasonable request.

Acknowledgement

This study was supported by the Swedish Childhood Cancer Foundation (grant no. MT2016-0016), the Swedish innovation agency Vinova (grant no. 2017-01247), and the Swedish Research Council (VR) (grant no. 2018-04375). We also thank Swedish Medtech4Health AIDA for giving us access to their Nvidia DGX-2 server.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejomp.2021.03.013>.

References

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68:394–424. <https://doi.org/10.3322/caac.21492>.
- Wu GX, Raz DJ. Lung Cancer Screening, Springer. Cham 2016;1–23. https://doi.org/10.1007/978-3-319-40389-2_1.
- The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N Engl J Med* 2011;409(365): 395. <https://doi.org/10.1056/NEJMOA1102873>.
- Austin JH, Müller NL, Friedman PJ, Hansell DM, Naidich DP, Remy-Jardin M, et al. Glossary of terms for CT of the lungs: recommendations of the Nomenclature Committee of the Fleischner Society. <http://DxDoiOrg/101148/Radiology20028685321> 1996. <https://doi.org/10.1148/RADIOLOGY.200.2.8685321>.
- Liu X, Hou F, Qin H, Hao A. Multi-view multi-scale CNNs for lung nodule type classification from CT images. *Pattern Recognit* 2018;77:262–75. <https://doi.org/10.1016/j.patcog.2017.12.022>.
- Loverdos K, Fotiadis A, Kontogianni C, Iliopoulou M, Gaga M. Lung nodules: A comprehensive review on current approach and management. *Ann Thorac Med* 2019;14:226–38. <https://doi.org/10.4103/atm.ATM.110.19>.
- Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network. *IEEE Trans Neural Networks Learn Syst* 2019;30:3484–95. <https://doi.org/10.1109/TNNLS.2019.2892409>.
- American college of Radiology. Lung CT screening reporting and data system (Lung-RADS) 2014. <https://www.acr.org/Clinical-Resources/Reporting-and-Data-a-Systems/Lung-Rads>.
- MacMahon H, Austin JHM, Gamsu G, Herold CJ, Jett JR, Naidich DP, et al. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the fleischner society. *Radiology* 2005;237:395–400. <https://doi.org/10.1148/radiol.2372041887>.
- Yu-Jen Chen Y-J, Hua K-L, Hsu C-H, Cheng W-H, Hidayati SC. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *Onco Targets Ther* 2015;8:2015. <https://doi.org/10.2147/OTT.S80733>.
- 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 Recognit* 2017;61:663–73. <https://doi.org/10.1016/j.patcog.2016.05.029>.
- Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 1, IEEE; n.d., p. 886–93. <https://doi.org/10.1109/CVPR.2005.177>.
- Ojala T, Pietikainen M, Maenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 2002;24:971–87. <https://doi.org/10.1109/TPAMI.2002.1017623>.
- Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014;5. <https://doi.org/10.1038/ncomms5006>.
- Chen C-H, Chang C-K, Tu C-Y, Liao W-C, Wu B-R, Chou K-T, et al. Radiomic features analysis in computed tomography images of lung nodule classification. *PLoS ONE* 2018;13:e0192002. <https://doi.org/10.1371/journal.pone.0192002>.
- Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, et al. Radiomics and radiogenomics in lung cancer: A review for the clinician. *Lung Cancer* 2018; 115:34–41. <https://doi.org/10.1016/j.lungcan.2017.10.015>.
- Wu W, Pierce LA, Zhang Y, Pipavath SNJ, Randolph TW, Lastwika KJ, et al. Comparison of prediction models with radiological semantic features and radiomics in lung cancer diagnosis of the pulmonary nodules: a case-control study. *Eur Radiol* 2019;29:6100–8. <https://doi.org/10.1007/s00330-019-06213-9>.
- Astaraki M, Wang C, Buizza G, Toma-Dasu I, Lazzaroni M, Smedby Ö. Early survival prediction in non-small cell lung cancer from PET/CT images using an intra-tumor partitioning method. *Phys Med* 2019;60:58–65. <https://doi.org/10.1016/j.ejomp.2019.03.024>.
- Han F, Wang H, Zhang G, Han H, Song B, Li L, et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *J Digit Imaging* 2015;28:99–115. <https://doi.org/10.1007/s10278-014-9718-8>.
- Lee MC, Borczyk L, Sungur-Stasik K, Cann AD, Borczuk AC, Kawut SM, et al. Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. *Artif Intell Med* 2010;50: 43–53. <https://doi.org/10.1016/j.artmed.2010.04.011>.
- Buty M, Xu Z, Gao M, Bagci U, Wu A, Mollura DJ. Characterization of Lung Nodule Malignancy Using Hybrid Shape and Appearance Features, Springer, Cham; 2016, p. 662–70. https://doi.org/10.1007/978-3-319-46720-7_77.
- Bonavita I, Rafael-Palou X, Ceresa M, Piella G, Ribas V, González Ballester MA. Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. *Comput Methods Programs Biomed* 2020;185:105172. <https://doi.org/10.1016/j.cmpb.2019.105172>.
- Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
- Shen S, Han SX, Aberle DR, Bui AA, Hsu W. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Syst Appl* 2019;128:84–95. <https://doi.org/10.1016/j.eswa.2019.01.048>.
- Hussein S, Cao K, Song Q, Bagci U. Risk Stratification of Lung Nodules Using 3D CNN-Based Multi-task Learning, Springer, Cham; 2017, p. 249–60. https://doi.org/10.1007/978-3-319-59050-9_20.
- Zhu W, Liu C, Fan W, Xie X. DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification. 2018 IEEE Winter Conf. Appl. Comput. Vis., IEEE; 2018, p. 673–81. <https://doi.org/10.1109/WACV.2018.000079>.
- Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, van Riel SJ, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Trans Med Imaging* 2016;35:1160–9. <https://doi.org/10.1109/TMI.2016.2536809>.
- Xie Y, Xia Y, Zhang J, Song Y, Feng D, Fulham M, et al. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. *IEEE Trans Med Imaging* 2019;38:991–1004. <https://doi.org/10.1109/TMI.2018.2876510>.
- Lei Y, Tian Y, Shan H, Zhang J, Wang G, Kalra MK. Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. *Med Image Anal* 2020;60:101628. <https://doi.org/10.1016/j.media.2019.101628>.
- Xie Y, Zhang J, Xia Y. Semi-supervised adversarial model for benign-malignant lung nodule classification on chest CT. *Med Image Anal* 2019;57:237–48. <https://doi.org/10.1016/j.media.2019.07.004>.
- Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial Autoencoders. *ArXiv:1511.05644* 2015.
- Wu B, Zhou Z, Wang J, Wang Y. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In: 2018 IEEE 15th Int. Symp. Biomed. Imaging (ISBI 2018), IEEE; 2018, p. 1109–13. <https://doi.org/10.1109/ISBI.2018.8363765>.
- Kumar D, Wong A, Clausi DA. Lung Nodule Classification Using Deep Features in CT Images. In: 2015 12th Conf. Comput. Robot Vis., IEEE; 2015, p. 133–8. <https://doi.org/10.1109/CRV.2015.25>.
- Q.Z. Song L, Zhao X.K. Luo of healthcare Engineering XCD-J, undefined, Using deep learning for classification of lung nodules on computed tomography images J Healthc Eng 2017 2017 2017 10.1155/2017/8314740.
- Fakoor R, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. In: Proc. ICML Work. Role Mach. Learn. Transform. Healthc.; 2013.
- Rasmus A, Berglund M, Honkala M, Valpola H, Raiko T. Semi-supervised Learning with Ladder Networks 2015:3546–54.
- Xie Y, Zhang J, Xia Y, Fulham M, Zhang Y. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Inf Fusion* 2018;42:102–10. <https://doi.org/10.1016/j.inffus.2017.10.005>.
- Causey JL, Zhang J, Ma S, Jiang B, Qualls JA, Polite DG, et al. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci Rep* 2018;8: 9286. <https://doi.org/10.1038/s41598-018-27569-w>.
- 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. *Med Phys* 2011;38:915–31. <https://doi.org/10.1118/1.3528204>.
- Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556* 2015.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recognit.; 2016.
- Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017 2016;2017-January:2261–9.
- Tan M, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks. In: 36th Int Conf Mach Learn ICLR 2019 2019;2019-June:10691–700.
- Kingma DP, Welling M. Auto-Encoding Variational Bayes. *ArXiv:1312.6114* 2014.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321–57. <https://doi.org/10.1613/jair.953>.
- Kaggle, “Data Science Bowl” 2017. <https://www.kaggle.com/c/data-science-bowl-2017>.
- Software Toolkit for Medical Image Analysis. <http://mialab.org/>.
- Dong C, Xue T, Wang C. The feature representation ability of variational autoencoder. In: Proc. - 2018 IEEE 3rd Int. Conf. Data Sci. Cyberspace, DSC 2018,

- Institute of Electrical and Electronics Engineers Inc.; 2018, p. 680–4. <https://doi.org/10.1109/DSC.2018.00108>.
- [49] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837. <https://doi.org/10.2307/2531595>.
- [50] Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. vol. 12. 2011.
- [51] Abadi M, Agarwal A, Bahram P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems 2015.
- [52] Yang J, Fang R, Ni B, Li Y, Xu Y, Li L. Probabilistic Radiomics: Ambiguous Diagnosis with Controllable Shape Analysis, Springer, Cham; 2019, p. 658–66. https://doi.org/10.1007/978-3-030-32226-7_73.
- [53] Nishio M, Nagashima C. Computer-aided diagnosis for lung cancer: usefulness of nodule heterogeneity. *Acad Radiol* 2017;24:328–36. <https://doi.org/10.1016/J.ACRA.2016.11.007>.
- [54] Lee H, Hong H, Seong J, Kim JS, Kim J. Treatment Response Prediction of Hepatocellular Carcinoma Patients from Abdominal CT Images with Deep Convolutional Neural Networks, Springer, Cham; 2019, p. 168–76. https://doi.org/10.1007/978-3-030-32281-6_18.
- [55] Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale Convolutional Neural Networks for Lung Nodule Classification, Springer, Cham; 2015, p. 588–99. https://doi.org/10.1007/978-3-319-19992-4_46.
- [56] Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network 2017. <https://doi.org/10.1109/TNNLS.2019.2892409>.