



# A multi-task CNN approach for lung nodule malignancy classification and characterization

Sónia Marques<sup>a</sup>, Filippo Schiavo<sup>a</sup>, Carlos A. Ferreira<sup>a,b,\*</sup>, João Pedrosa<sup>a</sup>, António Cunha<sup>a,c</sup>, Aurélio Campilho<sup>a,b</sup>

<sup>a</sup> INESC TEC, Porto, Portugal

<sup>b</sup> Faculdade de Engenharia da Universidade do Porto, Portugal

<sup>c</sup> Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal

## ARTICLE INFO

### Keywords:

Lung cancer  
Malignancy  
Multitasking classification  
Convolutional neural networks  
Deep learning

## ABSTRACT

Lung cancer is the type of cancer with highest mortality worldwide. Low-dose computerized tomography is the main tool used for lung cancer screening in clinical practice, allowing the visualization of lung nodules and the assessment of their malignancy. However, this evaluation is a complex task and subject to inter-observer variability, which has fueled the need for computer-aided diagnosis systems for lung nodule malignancy classification. While promising results have been obtained with automatic methods, it is often not straightforward to determine which features a given model is basing its decisions on and this lack of explainability can be a significant stumbling block in guaranteeing the adoption of automatic systems in clinical scenarios. Though visual malignancy assessment has a subjective component, radiologists strongly base their decision on nodule features such as nodule spiculation and texture, and a malignancy classification model should thus follow the same rationale. As such, this study focuses on the characterization of lung nodules as a means for the classification of nodules in terms of malignancy. For this purpose, different model architectures for nodule characterization are proposed and compared, with the final goal of malignancy classification. It is shown that models that combine direct malignancy prediction with specific branches for nodule characterization have a better performance than the remaining models, achieving an Area Under the Curve of 0.783. The most relevant features for malignancy classification according to the model were lobulation, spiculation and texture, which is found to be in line with current clinical practice.

## 1. Introduction

Lung cancer, among all kinds of cancers, has the highest mortality worldwide (Siegel, Miller, & Jemal, 2019). However, it has been shown that early-stage diagnosis of the disease significantly improves five-year survival rate (Knight et al., 2017), stressing the importance of implementing an efficient screening.

In clinical practice, low-dose computerized tomography (CT) is used by radiologists to find nodules, infer their characteristics and, ultimately their malignancy (Fig. 1). However, these tasks are highly complex and present significant observer variability. The prediction of malignancy from CT images is a particularly complex and subjective task, as radiologists typically rely on well-known nodule features to infer whether a nodule might be malignant or benign. Nodule malignancy classification is indeed a key step in the lung cancer screening process, as it allows to establish a diagnosis and determine patient follow-up.

Among other tasks in the lung cancer screening pipeline, computer-aided diagnosis (CAD) systems have been proposed as a tool to assist radiologists in nodule characterization and prediction of malignancy (Goncalves, Novo, Cunha, & Campilho, 2018). The large availability of labeled data (CT images with marked, segmented, characterized and classified nodules by experts in a methodological way), together with the growth of computational power, has given rise to the use of deep learning techniques with increasing performance. Among the available public datasets, the most widely known is the LIDC-IDRI (Armato et al., 2011), which contains 1018 CT scans, each annotated by four radiologists and comprising nodule location, segmentation, characterization and malignancy. Nodules with diameter  $\geq 3$  mm are characterized in terms of subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation and texture as well as malignancy, as described

\* Corresponding author at: INESC TEC, Porto, Portugal.

E-mail addresses: [soniamarques2@gmail.com](mailto:soniamarques2@gmail.com) (S. Marques), [filipposchiavo1994@gmail.com](mailto:filipposchiavo1994@gmail.com) (F. Schiavo), [carlos.a.ferreira@inesctec.pt](mailto:carlos.a.ferreira@inesctec.pt) (C.A. Ferreira), [joao.m.pedrosa@inesctec.pt](mailto:joao.m.pedrosa@inesctec.pt) (J. Pedrosa), [acunha@utad.pt](mailto:acunha@utad.pt) (A. Cunha), [campilho@fe.up.pt](mailto:campilho@fe.up.pt) (A. Campilho).

<https://doi.org/10.1016/j.eswa.2021.115469>

Received 28 September 2020; Received in revised form 23 April 2021; Accepted 21 June 2021

Available online 25 June 2021

0957-4174/© 2021 Elsevier Ltd. All rights reserved.

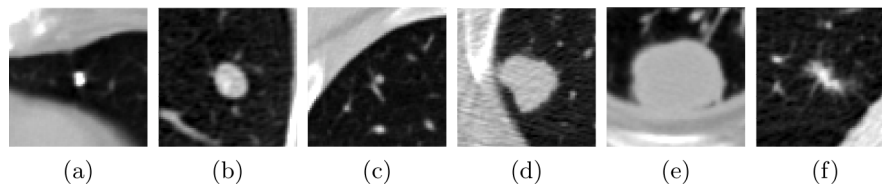


Fig. 1. Examples of benign (a–c) and malignant (d–f) nodules. Images show the axial slice crossing the nodule centroid ( $44.8 \times 44.8$  mm).

in Table 1. Further details on the nodule feature classes can be found in (McNitt-Gray et al., 2007).

Although nodules are the main sign of lung cancer, other state-of-the-art studies have used the entire slice for automatic classification. If, on the one hand, it is true that lung cancer can also induce other pulmonary changes, on the other hand, nodules may have little representation in the extracted features due to their smaller size and other features common to other lung diseases may be considered instead. In Nanglia, Kumar, Mahajan, Singh, and Rathee (2020), after extracting features within the lung region, a hybrid classification with supported vector machines (SVM) (Cortes & Vapnik, 1995) and neural networks, led to an overall classification accuracy of 98.08% in a set of cancer and non-cancer slices. Khan et al. (2020), after a contrast enhancement, used different algorithms to extract features concerning texture, invariant points and shape. Feature selection was then performed through weighted neighborhood component and classified by ensemble baggage tree, coming an accuracy of 99.4% per slice. Finally, Toğaçar, Ergen, and Cömert (2020), obtained an accuracy of 99.51% with a KNN classifier using a selection of features from the last fully connected layer of an AlexNet and achieved the most efficient classification accuracy at 98.74%.

For the classification of pulmonary nodules, to the best of our knowledge, Hua, Hsu, Hidayati, Cheng, and Chen (2015), were the first to describe the application of deep learning. These authors developed a deep belief network (DBN) (Hinton, 2009) and a convolutional neural network (CNN) for binary classification of nodule malignancy based on LIDC-IDRI. The DBN and the CNN-based methods achieved a sensitivity of 73.4% and 73.3%, and a specificity of 82.2% and 78.7%, respectively.

Hancock and Magnan (2016), presented a study on the relevance of the nodule features in LIDC-IDRI for the prediction of malignancy. Authors compared logistic regression (a linear method) and random forests (a non-linear method), concluding that the non-linear method produces more accurate results and then showed that spiculation, lobulation, subtlety, and calcification were the most relevant features for malignancy prediction.

Dai, Yan, Zheng, and Song (2018), implemented a multi-output 3D CNN based on 3D-ResNet-50 (He, Zhang, Ren, & Sun, 2016) and 3D-DenseNet-40 (Huang, Liu, van der Maaten, & Weinberger, 2017). In brief, the proposed architecture received as input CT images (a 3D cuboid around each nodule) and predicted malignancy as well as nodule features. The authors conclude that the inclusion of the characterization task in parallel to the malignancy prediction improved the results and that the features with the most impact on the results were spiculation, lobulation and calcification. An accuracy higher than 83% for all nodule features and of 91.47% for nodule malignancy classification was obtained LIDC-IDRI.

Shen, Han, Aberle, Bui, and Hsu (2019), presented a hierarchical CNN which received as input CT images (a 3D cuboid around each nodule) and predicted malignancy as well as five nodule features (sphericity, margin, subtlety, texture and calcification). In contrast to Dai et al. (2018), where characterization and malignancy prediction are parallel tasks, Shen et al. (2019), use intermediate features in the nodule characterization network branches for malignancy prediction. Authors claim that, given the association of nodule features with malignancy, the shared network modules and learnt nodule characterization

aid the network in learning generalizable features that are superior for malignancy prediction and show an improved performance when compared to a 3D CNN of similar architecture without intermediate tasks.

Liu, Dou, Chen, Qin, and Heng (2019), proposed a hierarchical architecture as in the work of Shen et al. (2019), but extended it for the prediction of all LIDC-IDRI nodule features. However, whereas Shen et al. performed binarization of the nodule feature and malignancy classes for classification, Liu et al. maintained the full range of nodule features classes (as shown in Table 1), using a regression module for nodule characterization. Malignancy is predicted by using intermediate features used in the nodule characterization task.

More recently, Bonavita et al. (2020) used an integrated approach that included nodule detection followed by malignancy classification with a CNN with three convolutional blocks and one dense block, obtaining a precision and recall of 83%.

While the link between nodule features and malignancy is well-established and promising results have been obtained combining malignancy prediction and nodule characterization tasks, the use of nodule features as support tasks for malignancy prediction is still an open problem. The fact that previous approaches use nodule characterization as a parallel or intermediate task, based on the shared use of intermediate layers, mean that a connection between the predicted malignancy and nodule features is not ensured, resulting in a less interpretable model. As such, in this study, we propose a method for malignancy prediction of lung nodules, which takes into account the prediction of nodule characterization, increasing model interpretability. Furthermore, a dedicated branch for the extraction of image features not described by the traditional nodule features is proposed, which gives the model the ability to learn features beyond those traditionally used by radiologists.

## 2. Methodology

### 2.1. Model architecture

With the two goals of malignancy classification and nodule feature characterization, four different architectures were designed, as shown in Table 2.

Model A is a neural network which receives as input the nodule features as graded by the radiologists. Models B–D are convolutional neural networks (CNN) which receive as input three parallel axial CT slices around the nodule's center. All models give as output the likelihood of malignancy. Models C and D also give as output the prediction of seven of the eight nodule features in LIDC-IDRI. Internal structure was excluded as its classes are severely imbalanced and it does not contain relevant information for malignancy prediction. The rationale for each of these models is the following:

- **Model A:** It replicates the final task of radiologists, predicting nodule malignancy from the annotated nodule features. The fact that standard/conventional features associated with malignancy are used means that a good performance is expected and the malignancy can be reasonably explained by the nodule features. However, the fact that it depends on manual annotations is a disadvantage as it would no longer represent a fully automatic approach.

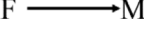



**Table 1**

Nodule features and meaning of 1–6 classes for each feature as defined on LIDC-IDRI annotations.

Nodule Feature	1	2	3	4	5	6
Internal structure	Soft tissue	Fluid	Fat	Air	–	–
Texture	Non-solid	...	Part solid	...	Solid	–
Subtlety	Extremel subtlety	Very subtle	Subtle	Relatively obvious	Obvious	–
Calcification	Popcorn	Laminated	Solid	Non-central	Central	Absent
Sphericity	Linear	...	Ovoid	...	Round	–
Margin	Poorly defined	...	...	...	Sharp	–
Lobulation	No lobulation	...	...	...	Marked	–
Spiculation	No spiculation	...	...	...	Marked	–
Malignancy	Highly unlikely	Moderately unlikely	Indeterminate	Moderately suspicious	Highly suspicious	–

**Table 2**

Overview of proposed model experiments.

Model	A	B	C	D
Input	Nodule features	CT image	CT image	CT image
Output	Malignancy	Malignancy	Malignancy and nodule features	Malignancy and nodules features
Architecture				

- **Model B:** It is designed for direct malignancy prediction from the CT image of a nodule. Although this model is expected to achieve very good classification performance, it is a black box as the image features extracted and used for malignancy classification are unknown when compared to model A.
- **Model C:** In order to merge the advantages of models A and B, model C was designed to predict the nodule features directly from the CT image and then use these features for malignancy prediction. In this way, a malignancy prediction from the image is obtained, as in model B, with the advantage that the features used would be known and are easily recognizable by radiologists as in model A, adding transparency and avoiding a black box behavior.
- **Model D:** Building on model C, model D includes a branch which is optimized solely to contribute to the prediction of malignancy, and not for the characterization of a feature as in model C. The rationale behind this additional branch is that there might be image features beyond those usually rated by radiologists that are related to malignancy. As such, model D has the freedom to extract these features and use them for malignancy prediction, along with the usual nodule features used by radiologists.

Fig. 2 shows the detailed architecture of the four proposed models. The architecture was adapted from previous work on the prediction of texture in pulmonary nodules (Ferreira, Aresta, Cunha, Mendonça, & Campilho, 2019; Ferreira, Cunha, Mendonça, & Campilho, 2018), which has shown promising results and is composed of a series of convolutional and dense blocks, with batch normalization (BatchNorm), (Ioffe & Szegedy, 2015), rectified linear unit activations (ReLU) (Glorot, Bordes, & Bengio, 2011) and dropout layers (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). In order to adapt this architecture for a multi-task approach, different branches of convolutional and dense blocks (denoted as  $B_F$  on Fig. 2) are considered. The first two convolutional blocks are shared among all branches to extract low-level features from the input image. Given that lobulation, margin, spiculation and subtlety are morphological features, weights are shared between these features for a more compact architecture. For sphericity, significantly worse results are reported in literature in comparison to other features (Shen et al., 2019) and it was empirically observed that having a separate branch for this feature and omitting the second convolutional block ( $B_{C3 \times 3 \ 56, 0.25}$ ) led to a significant improvement of performance. Softmax (Bishop, 2006) or sigmoid function (Han & Moraga, 1995) are applied to the output of the last dense layer of each feature to obtain the probability of each class (softmax for texture and sigmoid otherwise). The predicted features are then concatenated and fed to a fully connected neural network to obtain the probability of malignancy.

Note that each model is composed of different parts of the overall architecture shown in Fig. 2. Model A makes a direct prediction from the annotated nodules features, thus using only the fully connected neural network (yellow). Model B predicts malignancy from the CT image, thus using the first two convolutional blocks and the leftmost feature branch (blue) which outputs the malignancy probability. Model C predicts nodule features from the CT image, thus using the first two convolutional blocks and those related to feature prediction (green and blue). At the end, Model C uses the predicted nodule features for malignancy prediction through the fully connected neural network (yellow). Model D adds the branch dedicated only to image feature extraction (blue), thus using the full architecture.

## 2.2. Model training

Training was performed using 10-fold cross-validation. For each fold, the size of the validation set was chosen according to the minimum number of samples of the less represented class of spiculation, as it is one of the most imbalanced features.

Data augmentation was adopted to reduce overfitting through the use of random rotations, horizontal/vertical flips and zoom (Bloice, Roth, & Holzinger, 2019), applied online to each sample in the training set. This implies that, in a given epoch, a randomly transformed version of each training sample is observed once and that, in the following epochs, the random transformations applied to each sample are potentially different. The test set is not augmented in the automatic characterization of nodules.

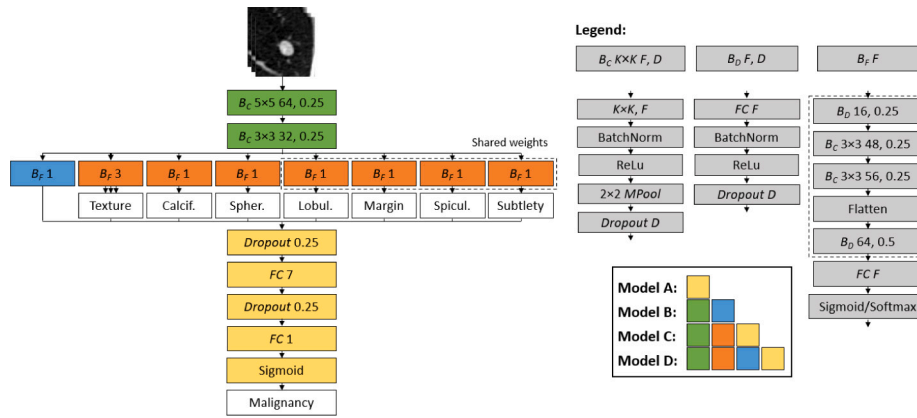
The loss function ( $L$ ) used during training was the weighted average of the malignancy and nodule feature losses defined as:

$$L = \sum_{f=1}^F \sigma_f L_f, \quad (1)$$

where  $F$  is the number of features (nodule features and malignancy) being optimized,  $L_f$  the loss function of feature  $f$  and  $\sigma_f$  the loss weight of that feature. Loss weights were determined empirically: 2.2 was used for sphericity, 0.7 for lobulation, margin, spiculation and subtlety and unitary loss weights for the remaining. Cross-entropy loss ( $L_{CE}$ ) was used for malignancy and all nodule features:

$$L_{CE} = - \sum_{c=0}^{M-1} \alpha_c y_{o,c} \log(p_{o,c}), \quad \alpha_c = \frac{N}{MN_c} \quad (2)$$

where  $M$  is the number of classes,  $\alpha_c$  is the class weight for class  $c$ ,  $y_{o,c}$  is the ground-truth label of class  $c$  for nodule  $o$  and  $p_{o,c}$  is the predicted probability of nodule  $o$  belonging to class  $c$ . Class weight  $\alpha_c$  is computed



**Fig. 2.** Schematic of the architectures implemented for nodule characterization and/or classification. Convolutional blocks  $B_c K \times K F, D$  correspond to a sequence of a convolutional layer ( $K \times K$ ) with a  $K \times K$  kernel and  $F$  output channels, BatchNorm and ReLU,  $2 \times 2$  max-pooling ( $2 \times 2$  MPool) and dropout (Dropout  $D$ ) with  $D$  probability. Dense blocks  $B_d F, D$  correspond to a sequence of a fully connected layer ( $FC F$ ) with  $F$  output channels, BatchNorm and ReLU and dropout (Dropout  $D$ ) with  $D$  probability. Feature blocks  $B_f F$  correspond to a sequence of dense and convolutional blocks, a fully connected layer of  $F$  output channels and a sigmoid/softmax layer.  $B_f$  weights (except for the last  $FC$ ) are shared on the lobulation, margin, spiculation and subtlety branches. Colors correspond to the elements of the architecture used in each of the models.

**Table 3**

Parameter settings for the performed experiments.

Parameters	Value
Max number of epochs	200
Patience	50
Batch size	32
Learning rate	0.002

for each feature and class according to the total data size  $N$  and the data size of class  $c$ ,  $N_c$ .

For models A and B training was done in a single phase, optimizing the prediction of malignancy. Models C and D were trained in two phases to improve the malignancy prediction as it corresponds to the main clinical goal. As such, a collective optimization of the prediction of all nodule features and malignancy was first performed, using the malignancy validation loss as an early stopping criteria. Secondly, the two first convolutional blocks and the branches that lead to feature prediction were frozen and optimization for the prediction of malignancy only was performed.

Table 3 presents the used parameter settings for the conducted experiments, selected empirically based on previous experience in various image recognition tasks. For the optimization of the models, the Nesterov Adam (Adaptive Moment Estimation) (Kingma & Ba, 2015) was used as optimizer, with a learning rate of 0.002 and a batch size of 32. A patience parameter of 50 epochs was used for the early stopping criterion.

### 2.3. Dataset

The main dataset used in this study is the LIDC-IDRI (Armato et al., 2011), which contains thoracic CT scans of 1010 patients, collected by seven academic centers and eight medical imaging companies. Each scan has been annotated by four radiologists (among twelve who participated in the study) in a two-phase process (blinded-read phase followed by an unblinded-read phase). The annotations comprise nodule location, segmentation, nodule features (as described in Table 1) and malignancy.

In total, LIDC-IDRI has 2668 pulmonary nodules with diameters greater than or equal to 3 mm. However, an unknown subset of 100 scans in the first 399 scans (containing 1053 nodules) presents a reversed labeling for lobulation and spiculation (LIDC-IDRI - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki, 2020), i.e. an unknown subset of radiologists were using a scale of 5–1 rather than 1–5. In order to address this issue, two measures were

undertaken: first, the nodules of the mentioned subset with a variance in the annotations of spiculation and lobulation greater than 1.7 or annotated by a single radiologist were excluded from the training and validation sets; second, the nodules of this subset were not included on the test set (regarding lobulation and spiculation). The first measure relates to the fact that for many nodules the problem of mislabeling probably regards a part of the four annotations and will thus remove cases where different scales have been used; the second decision avoids testing on nodules with possible wrong labels.

The images have  $64 \times 64$  pixels (pixelsize of 0.7 mm  $\times$  0.7 mm) spaced at 1.4 mm between each slice. CT slices are normalized from  $-1000$  and 400 Hounsfield Units to a range of  $[0, 1]$ . As detailed in Table 4, the average of annotations given by the radiologists for each nodule feature and malignancy was discretized. The threshold was chosen for each nodule feature so that the number of nodules in the minority is increased. In the particular case of malignancy, the rating 3 (indeterminate) on LIDC-IDRI was established as positive, not only because this was the minority class, but also because in a screening scenario indeterminate cases should be referred for further follow-up. For texture, the annotated ratings for nodules were considered into three classes commonly considered in radiology (solid, sub-solid and non-solid).

For external validation of the proposed architecture, the LNDb dataset (Pedrosa et al., 2021) was used, which contains 294 CT scans collected at the Centro Hospitalar e Universitário de São João (CHUSJ) in Porto, Portugal between 2016 and 2018. The acquisition and annotation protocols used for LNDb follow those of LIDC-IDRI with the exception that the annotations in LNDb were made in a single-blind fashion, i.e. radiologists performed a single annotation without reviewing annotations from other radiologists. Furthermore, annotations in LNDb comprise not only nodules  $\geq 3$  mm but also nodules  $< 3$  mm. As such, to ensure continuity with LIDC-IDRI in terms of nodule size and so that only reliable annotations were included, only nodules with diameter greater than 3 mm and annotated by more than two radiologists were selected from LNDb, resulting in a total of 77 nodules.

## 3. Experiments

### 3.1. Malignancy classification and feature characterization

All models were trained on LIDC-IDRI and tested on both the LIDC-IDRI test set and the LNDb dataset for each cross-validation fold.

Malignancy classification performance was evaluated using the receiver operating curve (ROC), i.e. true positive (TP) rate vs false positive (FP) rate, specifically through the area under the curve (AUC). Because



**Table 4**

Encoding LIDC-IDRI ordinal labels into classes, corresponding to the intervals defined in the table.

Nodule feature	Class		
	0	1	2
Texture	Non-solid [1, 2.3(3)[	Part-solid [2.3(3), 3.6(6)]	Solid ]3.6(6), 5]
Calcification	Present [1, 5.5[	Absent [5.5, 6]	–
Sphericity	Linear [1, 3]	Round ]3, 5]	–
Lobulation	Marked [1, 3[	None [3, 5]	–
Margin	Poorly defined [1, 3]	Sharp ]3, 5]	–
Spiculation	Marked [1, 3[	None [3, 5]	–
Subtlety	Subtle [1, 3]	Obvious ]3, 5]	–

the data is not extremely imbalanced, additional analysis through, for instance, partial AUC (Carrington et al., 2020) or precision–recall curves were not used, despite their relevance in similar problems.

Nodule feature characterization performance was evaluated in terms of accuracy for each of the features. Feature accuracy was computed for a prediction threshold of 0.50. For texture, which has 3 classes, the class with highest probability was considered to be the predicted class. A single output model identical to model B for each of the features (hereinafter referred to as model B') and a multi-task model for all features without malignancy prediction identical to the top section of model C (hereinafter referred to as model C') were also used for comparison to models C and D.

Paired two tailed t-test analysis was performed to compare model performance considering statistical significance at  $p < 0.05$  and  $p < 0.01$ .

### 3.2. Feature importance analysis

The importance of each of the nodule features used for malignancy classification in models A, C and D was then evaluated. This was done by randomizing each nodule feature values separately and observing the impact on AUC of the LIDC-IDRI test set. For model A, binary values were used for each feature, according to the model architecture. For models C and D, the input image was used to predict all nodule features and these were replaced by random values between 0 and 1, before feeding them back to the model for malignancy classification.

### 3.3. Patient diagnosis

Finally, diagnosis data available for a limited number of the nodules in LIDC-IDRI (LIDC-IDRI - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki, 2020) was used to test the reliability of radiologists' annotations when compared with the results of the proposed models for nodule classification. As there is no indication of the location of the nodules in the diagnosis data (only the patient's ID), only patients with a single nodule with diameter greater than 3 mm were considered, given that if the patient had more than one nodule, a correspondence to the nodule(s) with diagnosis data could not be made. Moreover, cases where the diagnosis method is 'unknown' were discarded. This led to a total of 29 nodules with diagnosis data.

The diagnosis data of the 29 nodules was taken as a second ground-truth and the performance of models B, C and D and of the radiologists' malignancy annotations were assessed. Model A was not included in this analysis since 28 of the 29 nodules were part of the 1053 nodules initially excluded for this model due to mislabeling (cf. Section 2.3).

### 3.4. Computational specifications

All experiments were performed using Keras 2.2.4 in Python 3.7 with TensorFlow as backend on a desktop with an Intel (R) Core TM i7-6700K CPU @ 4.00GHz×8, 32 GB RAM and an 8 GB Nvidia GeForce GTX 1080 GPU.

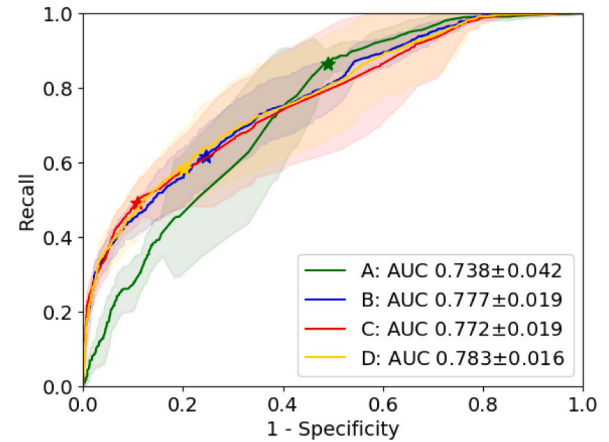


Fig. 3. ROC curves and AUC for malignancy predictions on LIDC-IDRI (star markers correspond to the threshold of 0.50 and shading represents standard deviation).

## 4. Results

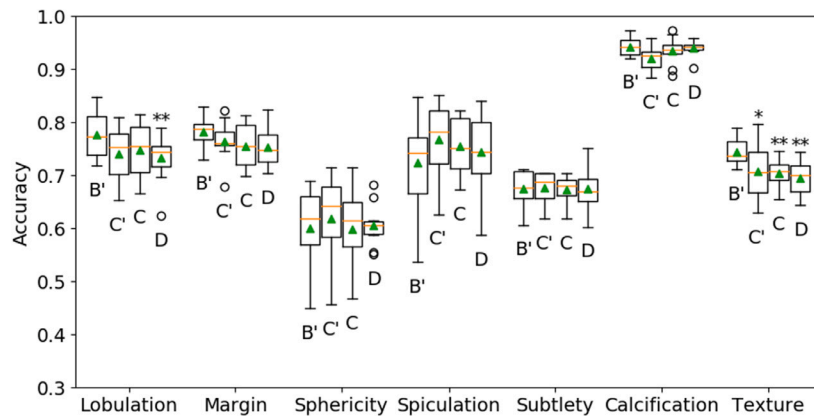
### 4.1. Malignancy classification and feature characterization

Fig. 3 shows the ROC curves and AUC of models A–D for malignancy classification on LIDC-IDRI. Statistical significant differences in AUC were found for every model at  $p < 0.01$  except between models A and C, where a statistically significant difference was found at  $p < 0.05$  and for models C and D ( $p < 0.15$ ). For a prediction threshold of 0.50, the results of malignancy classification on LIDC-IDRI were, for models A, B, C and D respectively  $0.68 \pm 0.04$ ,  $0.69 \pm 0.03$ ,  $0.70 \pm 0.02$  and  $0.69 \pm 0.02$  in terms of accuracy,  $0.86 \pm 0.04$ ,  $0.62 \pm 0.08$ ,  $0.49 \pm 0.05$  and  $0.58 \pm 0.10$  in terms of recall and  $0.51 \pm 0.05$ ,  $0.76 \pm 0.10$ ,  $0.89 \pm 0.04$  and  $0.80 \pm 0.09$  in terms of specificity.

The accuracy of the feature predictions of each model is shown in Fig. 4. Overall, the results for each feature are similar across models and statistically significant differences when compared to a single output network (B') were found only for lobulation (model D) and texture (models C, C' and D).

Examples of the results of model C and D are presented in Fig. 5, as these models allow a qualitative analysis of the explainability of the malignancy predictions. It was observed that predictions of the models tend to be more correct for images with well-defined nodules and high image quality. Juxtapleural nodules and low image quality tend to lead to a higher degree of incorrect predictions.

The malignancy classification performance on LNDb is shown in Fig. 6. Models B and D were those with better performance for LNDb. Statistical significant differences in AUC were found for every model at  $p < 0.01$  except between models A and C ( $p > 0.07$ ) and B and D ( $p > 0.54$ ). For a prediction threshold of 0.50, the results of malignancy classification on LNDb were, for models A, B, C and D respectively  $0.70 \pm 0.01$ ,  $0.66 \pm 0.11$ ,  $0.74 \pm 0.03$  and  $0.72 \pm 0.05$  in terms of accuracy,  $0.88 \pm 0.02$ ,  $0.85 \pm 0.05$ ,  $0.62 \pm 0.06$  and  $0.79 \pm 0.10$  in terms of recall and  $0.63 \pm 0.01$ ,  $0.60 \pm 0.15$ ,  $0.78 \pm 0.05$  and  $0.69 \pm 0.09$  in terms of specificity.



**Fig. 4.** Accuracy of feature characterization on LIDC-IDRI. Box limits and orange lines represent the first, second (median) and third quartiles of the data and green triangles represent the mean. Whisker ends represent the last data point within 1.5 the interquartile range of the first and third quartile respectively and circles represent outliers outside this range. \* and \*\* indicate a significant difference at respectively  $p < 0.05$  and  $p < 0.01$  to B' within each feature.

	Pred.	GT	Pred.	GT	Pred.	GT	Pred.	GT	Pred.	GT
Lobulation	0	0	1	1	0	0	<u>1</u>	0	<u>0</u>	1
Margin	1	1	0	0	<u>0</u>	<u>1</u>	0	0	<u>1</u>	0
Sphericity	1	1	1	1	0	<u>1</u>	0	0	1	1
Spiculation	0	0	1	1	0	0	<u>1</u>	0	<u>0</u>	<u>1</u>
Subtlety	1	1	1	1	0	0	1	1	0	0
Calcification	0	0	1	1	1	1	1	1	1	1
Texture	2	2	2	2	2	2	<u>2</u>	<u>1</u>	<u>2</u>	<u>1</u>
Malignancy	0	0	1	1	<u>1</u>	0	1	1	0	1

	Pred.	GT	Pred.	GT	Pred.	GT	Pred.	GT	Pred.	GT
Lobulation	0	0	1	1	0	0	<u>1</u>	0	0	1
Margin	1	1	1	1	1	1	0	0	<u>0</u>	<u>1</u>
Sphericity	<u>0</u>	<u>1</u>	1	1	1	1	<u>0</u>	<u>1</u>	<u>1</u>	0
Spiculation	0	0	1	1	0	0	<u>1</u>	<u>0</u>	<u>0</u>	<u>1</u>
Subtlety	0	0	1	1	1	1	1	1	1	1
Calcification	1	1	1	1	1	1	1	1	1	1
Texture	2	2	2	2	2	2	<u>1</u>	<u>0</u>	<u>1</u>	<u>2</u>
Malignancy	0	0	1	1	<u>0</u>	<u>1</u>	1	1	0	1

**Fig. 5.** Examples of feature characterization and malignancy classification for models C and D (first and second rows). Wrong predictions are shown in underline and bold. In the first and second columns are examples of correct predictions for the most important features (spiculation, lobulation and texture) and malignancy and the following are examples of variations of correct/incorrect characterization and classification predictions. GT stands for ground-truth and "Pred." stands for prediction.

The feature characterization results on LNDb are shown in Fig. 7. Statistically significant differences when compared to a single output network (B') were found for calcification (C', C and D) and texture (C and D). Compared to LIDC-IDRI, similar accuracies were found, except for calcification and texture for which respectively lower and higher accuracies were obtained.

#### 4.2. Feature importance analysis

Fig. 8 shows the importance of each nodule feature for models A, C and D. It can be seen that the most important features for malignancy classification are lobulation, spiculation, calcification and texture. For model D, the features extracted from the image also play a considerable role. Comparing the different models, it can be seen that model A depends more heavily on each feature, whereas models C and D are less dependent on each individual features. This is especially true for subtlety, calcification and texture which have significantly reduced importance.

#### 4.3. Patient diagnosis

Malignancy classification accuracy according to diagnosis confirmation on LIDC-IDRI is shown in Fig. 9. It can be seen that all models have a performance similar to the original radiologist annotations although model C has the best performance in terms of AUC, contrary to the results shown in Fig. 4.

### 5. Discussion

#### 5.1. Malignancy classification and feature characterization

Comparing the different models for malignancy classification, it can be seen that malignancy predictions improve when the image data is provided to the CNN architecture. While reasonable performance can be obtained directly from the nodule characteristics using model A, models B–D have access to information from the image, leading to

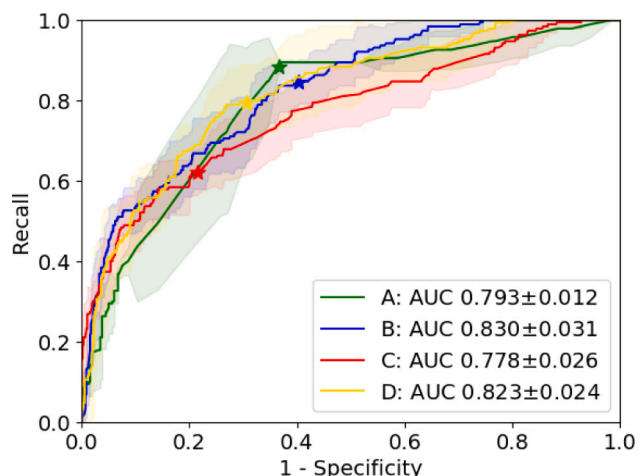


Fig. 6. ROC curves and AUC for malignancy predictions on LNDb (star markers correspond to the threshold of 0.50).

higher performance. Comparing specifically models A and C, which perform malignancy classification from the same nodule features, the

fact that model C has higher performance is probably due to the discretization of the input data used in model A, suggesting that continuous feature values contain valuable information for the prediction of malignancy.

While reasonable malignancy classification performance was obtained in this study, a lower accuracy was obtained when compared to state-of-the-art methods in literature. This can be due to several reasons related to the architecture design and the data used. The main goal of this study was to compare different architectures which provide nodule characterization as well as malignancy, when compared to direct methods. Moreover, the fact that an effort was made in this study to obtain a compact and low-parameter network can have negatively influenced the results. The fact that multiple 2D slices, rather than a fully 3D approach, were used can also contribute to lower performances, given that the network has access to limited data. Finally, the way that the data was treated may have played a role. Whereas previous studies have used radiologist annotations of the same nodule as separate entries (Shen et al., 2019), in this study radiologist annotations of the same nodule were averaged. While this approach increases the robustness of the features before discretization, it also leads to a significant decrease of the amount of data available for training, which may lead to decreased performance. In spite of the performance obtained, the methods proposed in this study, and specifically models C and D which predict malignancy associated with nodule characterization, are innovative and could be useful in clinical practice.

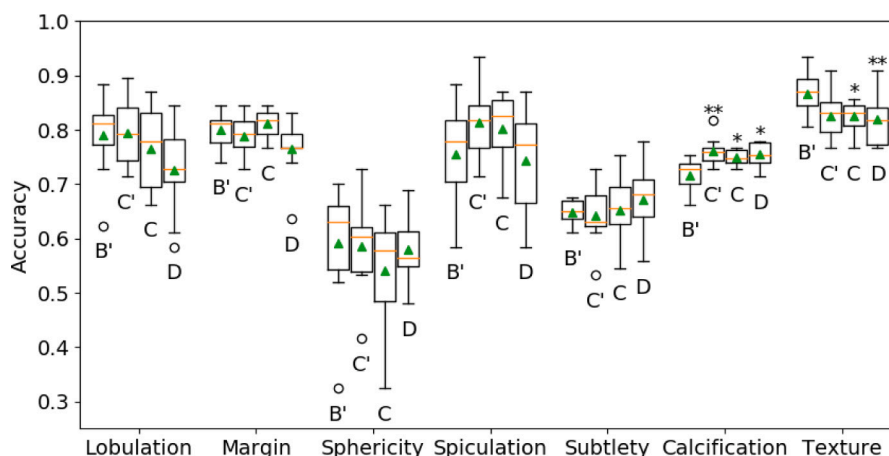


Fig. 7. Accuracy of feature characterization on LNDb. Box limits and orange lines represent the first, second (median) and third quartiles of the data and green triangles represent the mean. Whisker ends represent the last data point within 1.5 the interquartile range of the first and third quartile respectively and circles represent outliers outside this range. \* and \*\* indicate a significant difference at respectively  $p < 0.05$  and  $p < 0.01$  to B' within each feature.

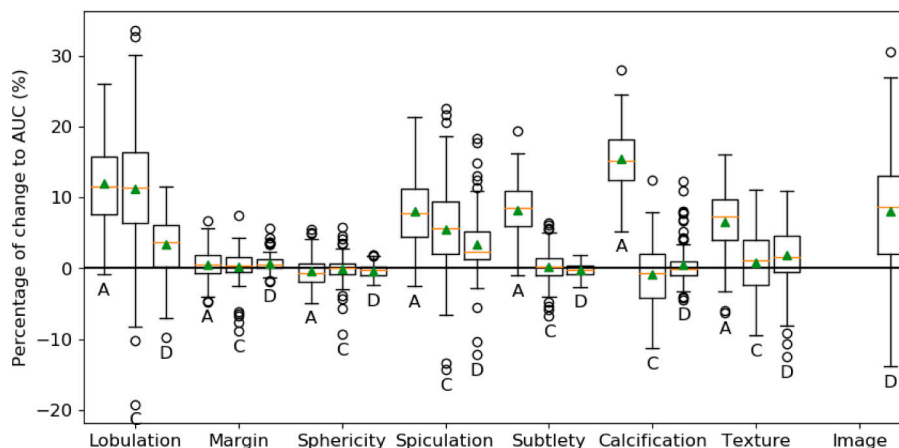


Fig. 8. Feature importance expressed in percentual change to AUC attributed to each feature and to the specialized branch for image features. Positive y-axis represents a decrease of AUC when respective feature values are randomized. Box limits and orange lines represent the first, second (median) and third quartiles of the data and green triangles represent the mean. Whisker ends represent the last data point within 1.5 the interquartile range of the first and third quartile respectively and circles represent outliers outside this range.

Regarding feature characterization, the fact that for all features except texture the performance of models C and D are not lower than the performance of a single output model (B') confirms that the simultaneous optimization of all features, rather than single-output optimization, is not detrimental. Furthermore, the fact that the performance of models C and D are not lower than the performance of a multitask model for nodule characterization (C'), assures that combining feature optimization with malignancy optimization does not jeopardize the nodule characterization results. Overall, reasonable accuracy was obtained for all features, though lower accuracy was obtained for sphericity and subtlety. This can be related to the difficulty and/or subjectiveness of these features given that sphericity and subtlety present the lowest observer agreement (0.35 and 0.40 respectively) and calcification presents the highest (0.92) (Pedrosa et al., 2019).

Regarding the results obtained for the LNDb dataset, these were similar to LIDC-IDRI, with image-based models having a superior performance, especially models B and D. However, the performance of model A was improved for LNDb, which may indicate that feature annotations by radiologists are more reproducible between datasets than the features extracted from the images acquired.

Regarding feature characterization on LNDb, the performance decrease observed for calcification is likely related to the underlying radiologist annotations in LNDb, where it has been shown that the observer agreement is significantly lower than in LIDC-IDRI (Pedrosa et al., 2019). The fact that a statistically significant higher performance was obtained for spiculation and calcification in a multi-task approach indicates that the multi-task approaches have generalized better. Given that the multi-task approaches use the same parameters for different features, the features learned are more likely to be representative of general aspects of the image, avoiding overfitting the data.

## 5.2. Feature importance analysis

With respect to the feature importance analysis, it can be seen that for model A, which takes as input the radiologist annotated nodule features, the most important features are in order of importance calcification, lobulation, spiculation, subtlety and texture. This is in agreement with previous studies using other architectures (Hancock & Magnan, 2016). It can also be seen that margin and sphericity have no impact on the AUC for any of the models. This indicates that either they have no/little predictive power for malignancy or they are highly redundant when combined with other nodule features, so that randomizing that information does not impact the overall malignancy classification.

For model C, which uses the same nodule features but predicted from image data, it can be seen that it relies less on each individual feature, as the change to AUC on the randomization of each feature is smaller than for model A. Furthermore, for subtlety, calcification and texture, the AUC does not significantly change. This can be due to the fact that model A relies on discretized values, whereas model C relies on floating-point values, which include a certain degree of uncertainty. The fact that the model was trained with uncertain rather than absolute values has made the model more robust to "errors" in the feature characterization step.

Regarding model D, it can be seen that randomizing the result from the image branch is detrimental to performance as well as traditional nodule features, namely lobulation and spiculation. Therefore, the network is taking nodule features into account as well as the image branch for malignancy classification. The fact that the image branch is important for malignancy classification indicates that there is information in the image that cannot be expressed by the traditional nodule features used by radiologists. These features are extracted by the model for malignancy classification, leading to a superior performance of model D when compared to C.

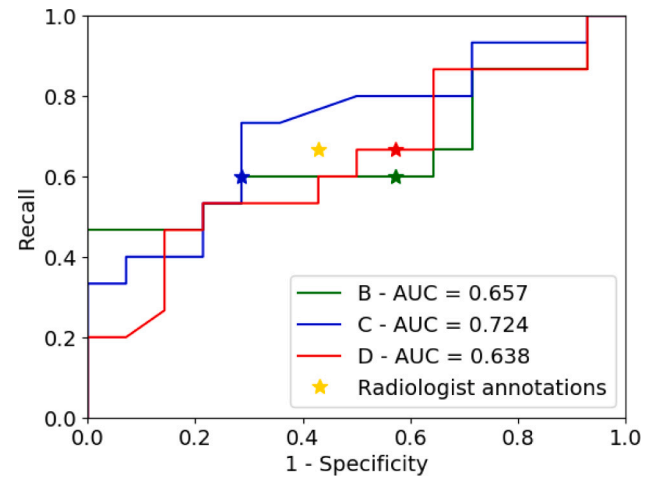


Fig. 9. ROC curves and AUC for malignancy predictions according to diagnosis confirmation on LIDC-IDRI (star markers correspond to the threshold of 0.50).

## 5.3. Patient diagnosis

Regarding the malignancy classification according to diagnosis confirmation, the amount of data used is too small to consider that the differences between models are significant. However, it is clear that, when compared to diagnosis confirmation, any of the models have a performance similar, if not superior, to radiologists.

## 6. Conclusions

This work shows an extensive performance comparison of different models with the goal of combining lung nodule characterization and classification, using a CNN approach, providing a study of the potential interpretability of malignancy classification.

Models B and D produced the best performance for nodule referral, achieving an AUC of 0.777 and 0.783 respectively on the LIDC-IDRI database. While both models perform malignancy prediction from the CT image of a nodule, model D has the additional advantage of combining direct malignancy prediction with specific branches for nodule characterization. This suggests that the malignancy and characteristics used by radiologists can be linked, allowing radiologists to better understand the malignancy predicted by the model and constituting a more useful CAD for clinical practice. As approached in the discussion, the use of full 3D could be an important step forward to increase the robustness of the classification, particularly for characteristics such as sphericity.

While promising results have been obtained, several limitations are present in this study that should be taken into account. While the main goal of this work was to predict malignancy in relation to the nodule features, and this goal has been achieved, the current model does not provide explainability of the decision at a nodule level. While the results show that lobulation and spiculation are the most important features for malignancy prediction, it is unknown whether, for an individual nodule, these features played a decisive role in the classification or not. This is, naturally, extremely important from a clinical point of view and will be the focus of future work.

Finally, data arrangement/preparation must be reconsidered. The binary discretization used can have a strong influence on the results and all classes or alternative discretizations could be used. This is especially true for malignancy for example, where a three-class rating could be used instead, given the high prevalence of class 3 (indeterminate malignancy). Furthermore, additional nodule features such as the diameter, volume or manual/automatic segmentation of the nodules could be used.



## CRediT authorship contribution statement

**Sónia Marques:** Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing, Visualization. **Filippo Schiavo:** Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Carlos A. Ferreira:** Conceptualization, Software, Writing - original draft, Writing - review & editing, Supervision. **João Pedrosa:** Software, Writing - original draft, Writing - review & editing, Visualization, Supervision. **António Cunha:** Writing - review & editing, Supervision. **Aurélio Campilho:** Writing - review & editing, Supervision.

## Declaration of competing interest

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

## Acknowledgments

This work was financed by the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness - COMPETE 2020 Programme and by National Funds through the Portuguese Funding agency, - FCT Fundação para a Ciência e a Tecnologia within project: PTDC/EEI-SII/6599/2014 (POCI-01-0145-FEDER-016673). Carlos A. Ferreira is funded by the FCT grant contract SFRH/BD/146437/2019.

## References

- Armato, S. 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. *Medical Physics*, 38(2), 915–931.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Bloice, M. D., Roth, P. M., & Holzinger, A. (2019). Biomedical image augmentation using augmentor. *Bioinformatics*, 35(21), 4522–4524.
- Bonavita, I., Rafael-Palou, X., Ceresa, M., Piella, G., Ribas, V., & Ballester, M. A. G. (2020). Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. *Computer Methods and Programs in Biomedicine*, 185, Article 105172.
- Carrington, A. M., Fieguth, P. W., Qazi, H., Holzinger, A., Chen, H. H., Mayr, F., et al. (2020). A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, 20(1), 1–12.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Dai, Y., Yan, S., Zheng, B., & Song, C. (2018). Incorporating automatically learned pulmonary nodule attributes into a convolutional neural network to improve accuracy of benign-malignant nodule classification. *Physics in Medicine & Biology*, 63(24).
- Ferreira, C. A., Aresta, G., Cunha, A., Mendonça, A. M., & Campilho, A. (2019). Wide residual network for Lung-Rads™ screening referral. In *2019 IEEE 6th portuguese meeting on bioengineering* (pp. 1–4). IEEE.
- Ferreira, C. A., Cunha, A., Mendonça, A. M., & Campilho, A. (2018). Convolutional neural network architectures for texture classification of pulmonary nodules. In *Iberoamerican congress on pattern recognition* (pp. 783–791). Springer.
- Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *14th international conference on artificial intelligence and statistics* (pp. 315–323).
- Goncalves, L., Novo, J., Cunha, A., & Campilho, A. (2018). Learning lung nodule malignancy likelihood from radiologist annotations or diagnosis data. *Journal of Medical and Biological Engineering*, 38(3), 424–442.
- Han, J., & Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks* (pp. 195–201). Springer.
- Hancock, M. C., & Magnan, J. F. (2016). Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: Probing the lung image database consortium dataset with two statistical learning methods. *Journal of Medical Imaging*, 3(4), Article 044504.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition*. IEEE.
- Hinton, G. (2009). Deep belief networks. *Scholarpedia*, 4(5), 5947.
- Hua, K.-L., Hsu, C.-H., Hidayati, S., Cheng, W.-H., & Chen, Y.-J. (2015). Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and Therapy*, 8, 2015–2022.
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition*. IEEE.
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
- Khan, M. A., Rubab, S., Kashif, A., Sharif, M. I., Muhammad, N., Shah, J. H., et al. (2020). Lungs cancer classification from CT images: An integrated design of contrast based classical features fusion and selection. *Pattern Recognition Letters*, 129, 77–85.
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Computing Research Repository (CoRR)*, abs/1412.6980.
- Knight, S. B., Crosbie, P. A., Balata, H., Chudziak, J., Hussell, T., & Dive, C. (2017). Progress and prospects of early detection in lung cancer. *Open Biology*, 7(9), Article 170070.
- LIDC-IDRI - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki (2020). <https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI>. Online (Accessed 02 January 2020).
- Liu, L., Dou, Q., Chen, H., Qin, J., & Heng, P.-A. (2019). Multi-task deep model with margin ranking loss for lung nodule analysis. *IEEE Transactions on Medical Imaging*.
- McNitt-Gray, M. F., Armato, S. G., Meyer, C. R., Reeves, A. P., McLennan, G., Pais, R. C., et al. (2007). The lung image database consortium (LIDC) data collection process for nodule detection and annotation. *Academic Radiology*, 14(12), 1464–1474.
- Nanglia, P., Kumar, S., Mahajan, A. N., Singh, P., & Rathee, D. (2020). A hybrid algorithm for lung cancer classification using SVM and neural networks. *ICT Express*.
- Pedrosa, J., Aresta, G., Ferreira, C., Atwal, G., Phoulady, H. A., Chen, X., et al. (2021). LNDb challenge on automatic lung cancer patient management. *Medical Image Analysis*, 70, Article 102027.
- Pedrosa, J., Aresta, G., Ferreira, C., Rodrigues, M., Leitão, P., Carvalho, A. S., et al. (2019). LNDb: A lung nodule database on computed tomography. arXiv:1911.08434.
- Shen, S., Han, S., Aberle, D., Bui, A., & Hsu, W. (2019). An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 128.
- Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1), 7–34.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929–1958.
- Toğaçar, M., Ergen, B., & Cömert, Z. (2020). Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks. *Biocybernetics and Biomedical Engineering*, 40(1), 23–39.