applied sciences

MDPI

Article

Pre-Training Autoencoder for Lung Nodule Malignancy Assessment Using CT Images

Francisco Silva [1,2], Tania Pereira [1], Julieta Frade [1,2], José Mendes [1,2], Claudia Freitas [3,4], Venceslau Hespanhol [3,4], José Luis Costa [4,5,6], António Cunha [1,7] and Hélder P. Oliveira [1,8,*]

1 INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, Porto 4200-465, Portugal; francisco.c.silva@inesctec.pt (F.S.); tania.pereira@inesctec.pt (T.P.); julietafrade97@gmail.com (J.F.); ee12293@fe.up.pt (J.M.); acunha@utad.pt (A.C.)
2 FEUP—Faculty of Engineering, University of Porto, Porto 4200-465, Portugal
3 CHUSJ—Department of Pulmonology, Centro Hospitalar e Universitário de São João, Porto 4200-319, Portugal; claudiaasfreitas@gmail.com (C.F.); hespanholv@gmail.com (V.H.);
4 FMUP—Faculty of Medicine, University of Porto, Porto 4200-319, Portugal; jcosta@ipatimup.pt
5 i3S—Institute for Research and Innovation in Health, University of Porto, Porto 4200-135, Portugal
6 IPATIMUP—Institute of Molecular Pathology and Immunology, University of Porto, Porto 4200-135, Portugal
7 UTAD—University of Trás-os-Montes and Alto Douro, Vila Real 5001-801, Portugal
8 FCUP—Faculty of Science, University of Porto, Porto 4169-007, Portugal
* Correspondence: helder.f.oliveira@inesctec.pt

Received: 28 September 2020; Accepted: 27 October 2020; Published: 5 November 2020

Abstract: Lung cancer late diagnosis has a large impact on the mortality rate numbers, leading to a very low five-year survival rate of 5%. This issue emphasises the importance of developing systems to support a diagnostic at earlier stages. Clinicians use Computed Tomography (CT) scans to assess the nodules and the likelihood of malignancy. Automatic solutions can help to make a faster and more accurate diagnosis, which is crucial for the early detection of lung cancer. Convolutional neural networks (CNN) based approaches have shown to provide a reliable feature extraction ability to detect the malignancy risk associated with pulmonary nodules. This type of approach requires a massive amount of data to model training, which usually represents a limitation in the biomedical field due to medical data privacy and security issues. Transfer learning (TL) methods have been widely explored in medical imaging applications, offering a solution to overcome problems related to the lack of training data publicly available. For the clinical annotations experts with a deep understanding of the complex physiological phenomena represented in the data are required, which represents a huge investment. In this direction, this work explored a TL method based on unsupervised learning achieved when training a Convolutional Autoencoder (CAE) using images in the same domain. For this, lung nodules from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) were extracted and used to train a CAE. Then, the encoder part was transferred, and the malignancy risk was assessed in a binary classification—benign and malignant lung nodules, achieving an Area Under the Curve (AUC) value of 0.936. To evaluate the reliability of this TL approach, the same architecture was trained from scratch and achieved an AUC value of 0.928. The results reported in this comparison suggested that the feature learning achieved when reconstructing the input with an encoder-decoder based architecture can be considered an useful knowledge that might allow overcoming labelling constraints.

Keywords: transfer learning; autoencoder; lung cancer; malignancy assessment

1. Introduction

Lung cancer is on the top of cancer-related mortality numbers worldwide [1,2]. Only 16% of lung cancer cases are diagnosed as local stage tumors. In these cases, patients have a five-year survival rate of more than 50%; however, when diagnosed in an advanced stage, the chances of a five-year survival decrease to 5%. Thus, achieving an earlier diagnosis is critical to increase survival rate, and systems able to provide screening support might play an important role. As a non-invasive method, computed tomography (CT) images have shown the ability to provide valuable information on tumor status, rising opportunities to the development of computer-aided diagnoses (CAD) systems able to provide an automatic assessment of lung nodules malignancy risk to help the clinical decision. Considering the use of qualitative data, factors like the high interobserver variability associated with the visual assessment of relevant characteristics, and the amount of radiological data to be analyzed makes the development of completely automatic systems a more attractive approach.

Several methods based on convolutional neural networks have been proposed to investigate the ability to distinguish between malignant and benign pulmonary nodules, taking advantage of the ability to directly detect relevant patterns with abstract and complex imaging manifestations [3]. Several previous works proposed deep learning-based solutions for lung nodule malignancy classification using the Lung Image Database Consortium image collection (LIDC-IDRI) [4,5], which is a public dataset of thoracic CT scans with expert annotations, and the most commonly used to develop AI-based solutions for lung cancer. Shen et al. [6] proposed a hierarchical learning framework to capture the nodule heterogeneity by utilizing a Convolutional neural network (CNN) to extract features and a random forest classifier for the final classification with the highest accuracy of 0.868. Lu et al. [7] obtained an accuracy of 0.919 using a CNN to extract the features and a support vector machine (SVM) for the final classification. Yan et al. [8] compared the performance obtained with 2D and 3D CNN implementations, achieving a mean area under the curve (AUC) of 0.937 for 2D analysis and an AUC of 0.947 using 3D inputs. Song et al. [9] proposed a comparison between a CNN, a deep neural network (DNN), and a stacked autoencoder (SAE) for the classification of benign or malignant lung nodules, and the CNN showed the highest performance, with an accuracy of 0.842. Yutong et al. [10] developed an algorithm that uses a deep convolutional neural network to automatically learn the feature representation of nodules on a 2D analysis, fuses this information with other more common features (shape, texture), and obtained an AUC of 0.966. A similar approach was developed by Causey et al. [11] that combines the deep learning CNN features with radiomics features as input in a random forest classifier and obtained an accuracy of 0.990. The success of deep learning-based feature extraction is due to the ability of not only taking the information of different conventional semantic features (such as shape or margins) or radiomic features (texture or histogram-based properties) but also taking into account abstract features, where deeper levels provide more complex and abstract knowledge [12,13].

Despite offering successful approaches, deep learning models require a large number of training data to be able to generalize over unseen images, and the lack of publicly available data is often a problem in the majority of medical applications. One strategy to overcome this issue is by using transfer learning (TL), a learning method that consists of applying a network already trained for a different task. By taking advantage of general patterns learned in the first layers, this technique reduces the number of trainable parameters alongside the necessary dataset size. Only a few and most recent works have explored this approach for this classification task. Lindsay et al. [14] used a pre-trained 3D-CNN on the LIDC-IDRI dataset to identify nodules in CT scans. Three new untrained layers were added to the existing pre-trained network, a private dataset of 796 patients who underwent CT-guided lung biopsy were used to retrain and test the approach. The biopsy results were used as ground truth labels removing the subjectivity of the annotations by the radiologists that are present in the LIDC-IDRI dataset. The TL of the initial layers of the networks with the retrain on the new dataset achieved an AUC of 0.70. Nóbegra et al. [15,16] proposed an investigation with multiple ImageNet [17,18] pre-trained feature extractors and different classifiers tested on the LIDC-IDRI dataset. The highest

AUC of 0.931 was achieved with the ResNet-50 [19] architecture and a SVM with a radial basis kernel as a classifier. Zhang et al. [20] explored a pulmonary nodule classification method based on the pre-trained ResNet for feature extraction and classification of the pulmonary nodules in an end-to-end manner that was evaluated on LIDC-IDRI and achieved an AUC of 0.912. Shi et al. [21] developed an approach based on a pre-trained VGG-16 model in ImageNet combined with an SVM classifier for a false-positive reduction in pulmonary nodule detection on CT slices from the LIDC-IDRI dataset and obtained an overall accuracy of 0.915.

The autoencoder is a dimensionality reduction method that adds the opportunity to extract features using unlabelled data, which allows overcoming one of the biggest limitations in the medical data. Autoencoders extract representative patterns from the images using the image reconstruction mechanism [22]. Kumar et al. [23] proposed to use an unsupervised denoising autoencoder to extract deep features from 2D lung nodule slices from the LIDC-IDRI dataset, achieving a classification mean accuracy value of 0.750, using a decision tree as a classifier. Cheng et al. [24] used a stacked denoising auto-encoder (SDAE) and constructed a pre-training architecture to use as network initialization for the latter supervised training, which achieved an AUC of 0.984 on the LIDC-IDRI dataset.

The work presented in this paper addressed a binary lung nodule malignancy classification by a TL approach based on a trained Convolutional Autoencoder (CAE), using the LIDC-IDRI dataset. Taking advantage of the unsupervised feature learning ability of the CAE while reconstructing nodule 2D images, the relevance of the learned patterns was explored to prevent the overfitting occurrence. Thus, a first experiment was conducted to train the CAE, and then a classification model was developed to distinguish between benign and malignant lung nodules in a 2D perspective. The major contribution of this work relies on three main points: simple architecture, by taking advantage of the computational resources lower consumption; no annotations required on pre-training, allowing to use large available datasets without the huge investment on labeling all data; pre-trained architecture, without the need of explicit design and selection of problem-oriented features that can be used for other tasks.

2. Materials and Methods

This section presents the dataset used in this work and describes the pipeline implemented to predict the malignancy of the lung nodules.

2.1. CT Image Database

The LIDC-IDRI [4,5] is a lung cancer screening dataset which comprises thoracic CT scans for a total of 1010 patients, alongside with annotated nodules belonging to one of three classes: (a) nodule $\geq$ 3 mm; (b) nodule < 3 mm or (c) non-nodule $\geq$ 3 mm, made during a two-phase annotation process by four experienced radiologists. Regarding data acquisition, slice thickness ranged from 0.6 to 5.0 mm, with X-ray current ranged from 40 to 627 mA (mean: 222.1 mA) at 120–140 kVp.

Data Preparation

To standardize the dataset, all CT scans were resampled. The pixel spacing was set to [1.00, 1.00, 1.00] mm and each CT dimension was calculated to match this new spacing, obtaining the resampled image by interpolation. Additionally, each pixel intensity value, measured in the Hounsfield Units (HU) scale, was normalized using the *min-max* normalization method, and values under -1000 HU, which corresponds to air's radiodensity value, were transformed into 0 and values above 400 HU, representing hard tissues like bones, were transformed into 1. A linear transformation was computed to map all values in the middle into the [0, 1] intended range.

To assess the nodule location, the final binary mask that included the pixels with at least 50% consensus of the provided contours was used, and an image with size ($80 \times 80 \times 80$) voxels centered on the nodule was extracted for each considered example in this study.

From the 7371 detected nodules, only 2669 were classified as larger than 3 mm. For these examples, this database also provides nodule contours marked by each radiologist, as well as quantitative labels

for a set of nodule related features. Although the feature learning task did not require any labels, this inclusion criterium was the one selected to facilitate the evaluation process. It was considered that a reconstructed image with a larger nodule would provide more clear information to evaluate the quality of the outputs by visualization.

Considering data inclusion in the classification task, the provided malignancy risk value assessed by the radiologists was analyzed. This value corresponded to an integer ranging from 1 to 5 with the following designations related to the malignancy degree: (1) *Highly Unlikely*, (2) *Moderately Unlikely*, (3) *Indeterminate*, (4) *Moderately Suspicious* and (5) *Highly Suspicious*. The interobserver agreement in rating malignancy was previously studied and a modest agreement was found [25,26]. For cases rated by two radiologists, they only agree in 43% of the cases. The percentage of agreement decreases for nodules annotated by more radiologists, with an agreement of 19% in the cases where the annotation was performed by three radiologists, and 12% of agreement in annotations from four radiologists. In addition, the analysis found in [5] regarding the percentage of nodules that were labelled by one, two, three or four radiologists, shows that almost 64% of the 2669 nodules were marked by a single clinician or by all of them, 29.1% and 34.8% respectively. When only assigned by one radiologist, the model only takes into account a single opinion, being susceptible to a possible error; on the other hand, if a nodule was labelled by all the four clinicians, the computed agreement rate of 12% tells that a lack of confidence in the resultant label would continue to exist. Having this analysis in mind and the fact that the subjectivity in the labelling process would not be eliminated if only the nodules annotated by all clinicians were included, this study includes the entire set of nodules marked by at least one radiologist.

To take into account the different annotations provided for the same example, each nodule's malignancy value was computed by averaging the provided annotations made by all radiologists, and a mean malignancy value $\leq$2.0 was considered as benign, and $\geq$4.0 as malignant, excluding the examples with resultant mean value outside the selected ranges. Although this database provides multi-class labeled nodules, this work aimed to perform a binary classification given the high inter-observer variability present in the available annotations. Considering the original range of values used to label each nodule, the arithmetic mean combined the different classifications with an equal contribution of each annotation [27]. This process helped to decrease the chances of using mislabelled information in the training or evaluation phases. Considering the explained inclusion criteria, only 1095 nodules were included in this classification task: 789 benign and 306 malignant.

The models implemented in this work were designed to receive two-dimensional inputs. Although this slice-level approach does not allow a complete nodule analysis, it increased the number of training examples by sampling different slices from the same nodule and also helped to achieve a better class balancing. Considering this, in the feature learning task, the slice over-sampling operation consisted of extracting middle slices from the axial, coronal and sagittal planes from the ($80 \times 80 \times 80$) nodule centered image, as illustrated in Figure 1. In the classification task, this operation included the extraction of four additional random slices to balance positive and negative classes in the training set, using the cube symmetry planes to obtain these extra nodule perspectives.

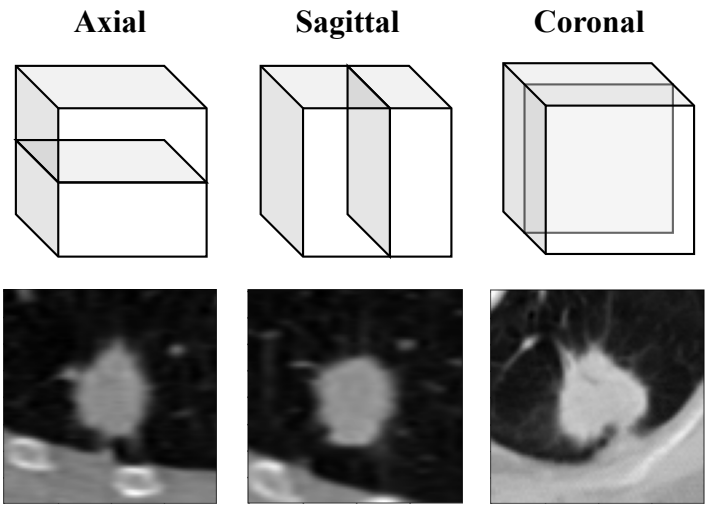

Figure 1. Lung nodule slice extraction example: middle slice from axial, sagittal and coronal planes from the same 3D nodule.

2.2. Methodology

The TL approach used in this work consisted of training a CAE to use the encoder layers as a feature extractor, followed by a classifier that uses those features to produce the final malignancy binary classification. Considering basic observation on CNN behavior, the first convolutional layers of the encoder learn generic features useful for many different tasks, but progressively, as the network goes deeper, more specific patterns are detected and more relevant information is learned regarding the target task [28]. In this approach, the relevance of the knowledge achieved while pre-training the encoder layers was explored in order to reduce the number of trainable parameters in the classification task. The pipeline proposed in this work is illustrated in Figure 2.

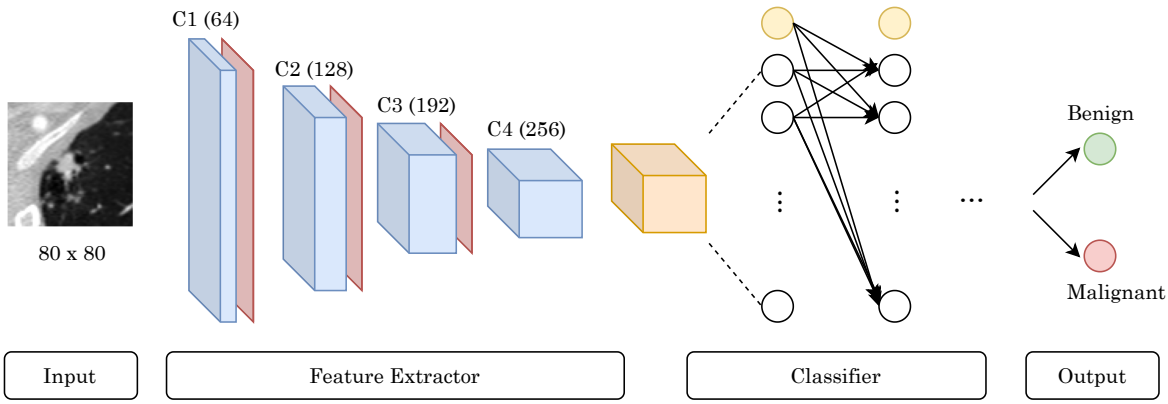

Figure 2. Pipeline developed for lung nodule malignancy classification composed by the feature extractor (CAE encoder) followed by a classifier.

In this classification task, a multi-stage training strategy was adopted, where initially only the classifier fully connected layers were trained. Given the fact that the encoder layers were pre-trained with images from the same domain, an initially good convergence was expected, but still far from a local minimum. Thus, to find the best classification performance, the convolutional layers were progressively unfrozen and retrained to detect more representative patterns related to nodule malignancy. By analyzing the learning curves, when learning stopped improving, a deeper convolutional layer was unfrozen, repeating this process until the overfitting occurrence. The major advantage of this layer-wise fine-tuning approach relies on the fact that immediately unfreezes the entire feature extractor at once, which was not expected to be a viable option given the small amount

of training data. Due to large similarities between source and target domains, it was also considered an unnecessary approach.

In training, validation, and testing phases, completely independent subsets of the LIDC-IDRI cohort were used to avoid the occurrence of images from the same nodule in more than one of those sets.

2.2.1. Feature Extraction

Extending the Autoencoder, a CAE enables a dimensionality reduction while preserving the original spatial representation, which is an absolutely critical characteristic when analyzing multi-dimensional data [29]. The overall structure of a CAE comprises two main phases: (1) in the encoding phase, a convolutional encoder transforms the input data into a lower-dimensional feature space, followed by a decoding phase (2), where the compressed representation passes through decoding layers to achieve the original input image reconstruction. Having the input as the target, the CAE learns the best set of features that enables the input reconstruction while eliminating the need for labeled data. The application of this dimensionality reduction technique lies in the idea that the knowledge achieved while training to reconstruct the input data might provide a useful weight initialization if one intends to use the encoder part for a different task in a transfer learning approach.

In the proposed CAE architecture, represented in Figure 3, the encoding phase is composed of four convolutional layers with 3×3 kernels, and an increasing number of filters as the network goes deeper. All the convolutional layers are followed by a Rectified Linear Unit (ReLU) activation function and a *max-pooling* layer to reduce the output feature map by half. Giving an input tensor of size $C \times H \times W$, passing through the encoding layers results in a feature map of 256 filters with size $\frac{H}{8} \times \frac{W}{8}$. To reconstruct the original input, three *max-unpooling* layers were employed to double the input size before the first three convolution blocks. The last convolutional layer is followed by a sigmoid activation, ensuring that all output pixels belong to the $[0, 1]$ range of values.

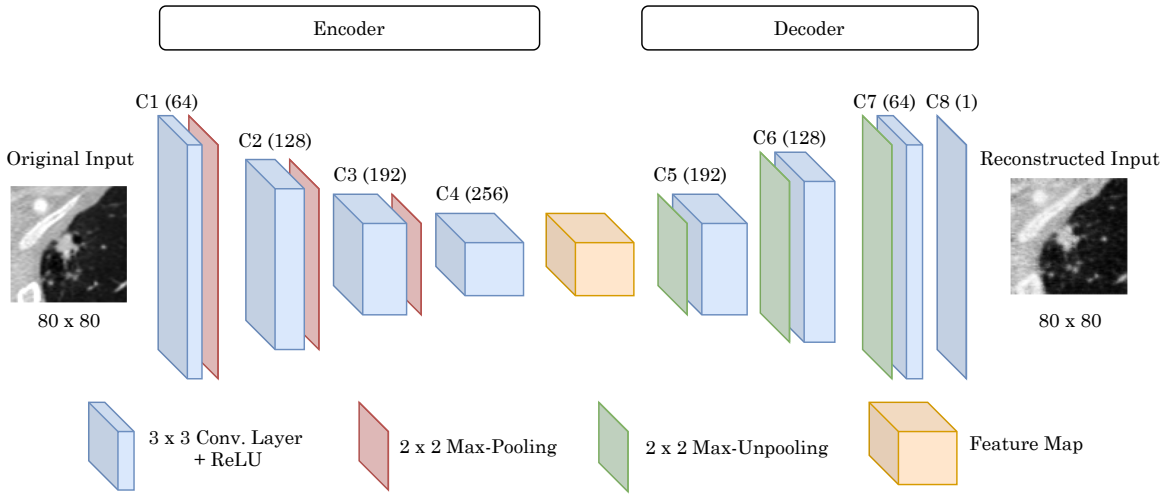

Figure 3. Proposed CAE architecture for lung nodule unsupervised feature learning by optimizing input reconstructions.

Table 1 depicts the considered values in the hyper-parameter search.

Table 1. Set of hyper-parameters values considered in the search for the CAE training.

Hyper-Parameter	Range Values
Learning Rate	0.0001, 0.001, 0.01, 0.1
Optimizer	SGD [1], Adam
Momentum	0.1, 0.5, 0.9

[1] Stochastic Gradient Descent.

2.2.2. Malignancy Classification

Given the end-to-end characteristic of this approach and the backpropagation based learning, a Multi-layer Perceptron (MLP) was used as a classifier to perform the intended predictions. An MLP is an artificial neural network composed of an input layer where all the input values are received, an output layer with a number of neurons depending on the classification task in hands, and in between a variable number of hidden layers. Since it is a fully-connected neural network, each neuron is connected to all neurons of the following layer. When approaching a classification task with TL techniques, the developed model will consist of the feature extraction pre-trained layers and a classifier stacked on top to be completely trained with the new target data (Figure 2). Each feature extracted by the convolutional block is used to feed the input fully-connected layer of the classifier. For regularization, a dropout layer was employed before this input layer to reduce the number of features taken into account at each training iteration, forcing the classifier to learn in a more robust manner. Additionally, the output of each neuron in the input and hidden layers passed through a ReLU activation, and a sigmoid activation was employed for the output neuron. The binary cross-entropy was used as the cost function to be minimized. The MLP is a backpropagation neural network, which works by approximating the non-linear relationship between the input and the output by adjusting the weight values internally. The ability of this neural network-based classifier to propagate the prediction error to the feature extractor layers allowed fine-tuning the higher-level layers of the convolutional encoder, where the most useful patterns are detected. The feature extractor is an encoder network pre-trained to reconstruct images in the same domain, not optimised to detect representative patterns related to nodule malignancy. To capture the most useful features for this classification task, the feature extractor needed to be fine-tuned, and this process is possible due to the backpropagation-based learning of MLP.

In this binary malignancy classification task, a more extensive search was employed to find the set of hyper-parameters that achieved the desired performance, with considered values depicted in Table 2.

Table 2. Set of hyper-parameters values used in the search for the malignancy classification model.

Hyper-Parameter	Range Values
Learning Rate	0.0001, 0.001, 0.01, 0.1
Batch-size	4, 8, 16, 32, 64
Momentum	0.1, 0.5, 0.9, 0.99
Weight decay	0.00001, 0.0001, 0.001, 0.01
Dropout	0.25, 0.5, 0.75
Hidden Layers	1, 2, 3
Hidden Neurons	32, 64, 128, 256, 512
Optimizer	SGD [1], Adam

[1] Stochastic Gradient Descent.

2.3. Performance Metrics

As common choice for learning, the Mean Squared Error (MSE) was used as the loss function, representing the averaged error of each output pixel value when compared to the same pixel in the original image [30,31]. Training stopped when no change in loss value was reported during 50 consecutive training iterations.

Considering the CAE training, the optimization criterium was based on the Mean Squared Error (MSE) value computed [32]. As a feature learning task, and even though a perfect reconstruction does not ensure the best set of learned features to be applied in the transfer learning approach under investigation, these experiments were conducted in order to optimize the input and its reconstruction similarities.

To evaluate the malignancy classification performance, the Receiver Operation Characteristic (ROC) curve was analyzed, as well as the AUC value. Additionally, Precision, Recall, and F-score metrics were also computed to measure the generalization ability of the developed models. To assess the reliability of the transfer learning approach adopted, the same architecture was also completely trained from scratch and evaluated.

3. Results

We evaluated the malignancy predictions in two training methods: using transfer learning and trained from the scratch. The results for the approach optimization and the classification performance achieved are presented in this section.

3.1. Hyper-Parameters Selection

The hyper-parameters selected for the CAE training in the nodule reconstruction task that achieved the best results were the following: mini-batches of four images with Stochastic Gradient Descent (SGD) as the optimizer, learning rate of 0.01, and momentum with the value of 0.9.

Considering the classification task, the hyper-parameters values that achieved the best performance on the test set are presented in Table 3. All the classification performance metrics were computed in a hold-out test set with 5-fold cross-validation applied to the training data, over five random train/test combinations to prevent some possible bias on results induced by a specific set of inputs.

Table 3. Hyper-parameters values that achieved best performance.

Hyper-Parameter	Value
Learning Rate	0.001
Batch-size	8
Momentum	0.9
Weight decay	0.0001
Dropout	0.5
Hidden Layers	1
Hidden Neurons	64
Optimizer	SGD [1]

[1] Stochastic Gradient Descent.

3.2. Malignancy Classification

The classifier training convergence was achieved after 200 iterations. After this first training stage, fine-tuning was applied by unfreezing the encoder's last convolutional layer; finally, after the second training convergence, gradients were updated for the last two convolutional layers for a final training. This strategy obtained the best results by preventing the model to overfit, which occurred when it was tried to add one more convolutional layer for retraining.

Mean values for each performance metric considered are presented in Table 4. By taking advantage of cross-validation data, the decision threshold was tuned for F-score maximization by the recall. This threshold optimization was employed by evaluating the precision-recall trade-off which, given the context, represents the cost of a missing malignant nodule (false negative) over a false suspicious of a benign tumor (false positive). Thus, as a missing malignant tumor should be a more penalized error, recall was maximized over precision in this classification task. With the TL approach, the mean value for

the classification threshold was 0.413 ± 0.101, which clearly illustrates the necessity of choosing a value under 0.5, the default threshold value, to allow to increase the confidence in non-malignant predictions.

Figure 4 shows the ROC curve for the binary classification between benign and malignant cases for our proposed method using the TL approach, with a maximum mean AUC of 0.936 ± 0.009. The small gray shading in the figure shows the consistency of the results and the independence of the performance with the subsets selected for training and testing, given the small variation on the AUC obtained.

Table 4. Lung nodule malignancy classification results.

	Performance Metrics (Mean ± Standard Deviation)			
Training Method	**AUC**	**Precision**	**Recall**	**F-Score**
Transfer Learning	0.936 ± 0.009	0.794 ± 0.026	0.848 ± 0.035	0.817 ± 0.020
Trained from scratch	0.928 ± 0.027	0.842 ± 0.035	0.789 ± 0.069	0.808 ± 0.022

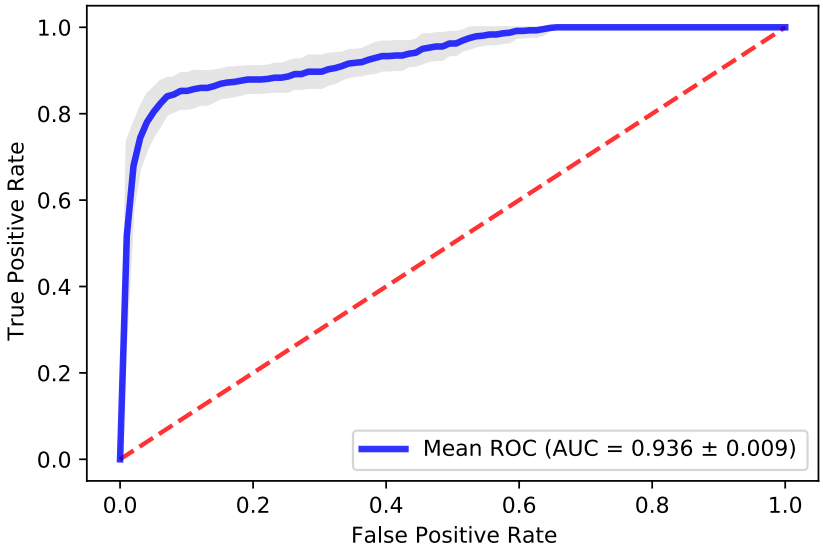

Figure 4. Averaged ROC curve for lung nodule malignancy classification using the transfer learning approach. The ROC curve is computed for each iteration, the arithmetic average is then calculated and represented by the blue line with a standard deviation represented by the gray shading area. The red dashed-line represents an at-chance classifier ROC curve.

To better understand the relationship between the success of a prediction and the output reconstruction of the correspondent nodule, Figure 5 shows one example for each case. Well and wrongly classified examples are depicted, alongside with the correspondent MSE value obtained by analysis of the CAE output. The reported MSE mean value for the successfully classified examples was 0.00162 ± 0.00088, in contrast with the value 0.00337 ± 0.00249, correspondent to wrong predictions. It was possible to verify that, in general, a correct prediction was associated with a reconstructed image with a lower pixel error, which might demonstrate the impact of the low-level features learning in the detection of more relevant patterns for classification.

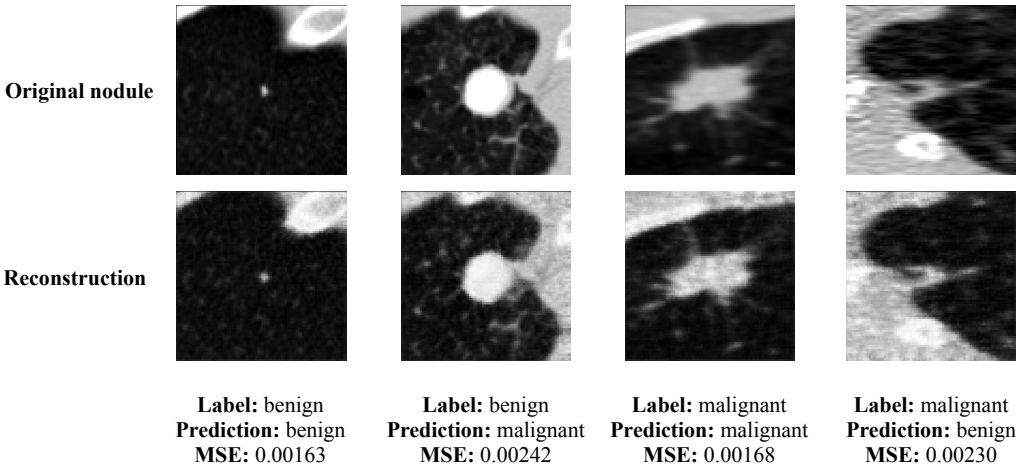

Figure 5. Testing examples, representing benign and malignant nodules, well and wrongly classified. The MSE value reported for each example represents the mean squared pixel error between the original image and its reconstruction.

4. Discussion

In the study, we proposed a deep structured algorithm to automatically extract features based on a convolutional autoencoder and an end-to-end learning classification network to predict the malignancy risk of nodules in CT images using TL techniques. A baseline experiment was also implemented, where the proposed architecture was trained from the scratch, in order to assess the capability to use the proposed TL strategy for this classification task. The results achieved suggest that the transfer learning approach was able to perform as well as the network trained from scratch, which means that the unsupervised learning ability of the proposed CAE represented a useful knowledge for the classification task, helping the network to avoid overfitting.

Considering the previous works, it is not possible to make a fair and direct comparison of the performance results, since the selection of the data and the criteria for final labelling of the nodules were different, which impacts the final performance. The results obtained in the current work did not overcome the performance obtained in some studies but achieve the same high level (above 0.9 of accuracy) using an approach that is not dependent on the massive label data and without needing feature engineering. In fact, the presented approach has several advantages that makes this contribution relevant in the medical image classification: (1) end-to-end approach (automatic feature extraction avoiding the *ad hoc* feature engineering); (2) unsupervised learning (allowing to use massive datasets without annotations to train the feature extractor) and (3) transfer learning (use the knowledge acquired with an unlabelled dataset to use in a different but related problem).

Additionally, studies using TL were usually based on pre-trained ImageNet models, since this is a massive annotated dataset. In a comparison with this more conventional transfer learning strategy, an important advantage provided by the proposed approach relies on the similarities between source and target domains, which is often a problem when applying a model trained on natural images in a medical imaging task like CT scans. This domain gap might compromise the feature extraction since the pre-trained model might not be able to detect the relevant features for the classification task, which might cause an impact on the required fine-tuning depth, alongside with the dataset size required for this deeper fine-tuning. The approach presented in this study allows the use of public CT datasets to train the feature extractor, ensuring that the model can capture the relevant features for the final problem.

One of the limitations of this work is the binary output characteristic. The classification performed by the clinicians considers multiple classes since the malignancy evaluation in a clinical environment is not limited by a binary benign/malignant result. Different classes might give additional and useful information to help with diagnosis. As future work, a multi-class solution will be developed. However,

a very clear annotated dataset must be labeled in order to decrease the label noise introduced by the inter-observer variability of the annotations. With a fine-grained classification, it is expected a higher disagreement between annotators, due to the higher difficult to distinguish between multiple classes. Several scientific societies recommended guidelines for nodule management [33]. The first assessment of the nodule is based on patient history and imaging studies. Further invasive investigation with biopsies is performed for the ones evaluated with a higher risk of malignancy. The fact that the malignancy label was based on subjective ratings by radiologists and not in an objective result such as a biopsy for all cases on the dataset, represents a limitation that will limit the performance that can be achieved by the developed models [14].

Some improvements might also be important to note to address other limitations in this work. Lung nodules are 3D elements, and with 2D or even 2.5D analysis, a large portion of useful information might be lost, which makes these perspectives sub-optimal ways of approaching this classification task. However, besides consuming more computational resources, a 3D approach does not allow a slice oversampling operation as employed in this study, which might be a problem given the lower amount of available training data, as well as the class imbalance naturally present. In the proposed feature learning task, the CAE is trained to minimize a cost function based on the pixel-wise error between the input and the correspondent reconstruction. The use of this loss might lead to blurred areas in the reconstructed images, meaning that the high-frequency components of the original image were not clearly learned by the encoder. These areas often correspond to edges or other detailed shapes, and clear learning of these low-level features might play an important role in the detection of relevant patterns related to the lung nodule malignancy [34].

Finally, the proposed approach is still limited on the explainable level since it is an end-to-end solution, which represents a black box for the clinicians. The importance of interpretability in machine learning is increasing due to the need to trust the final classification and understand the information used by the models for the prediction. The novel AI-based solutions should be transparent and understandable in order to generate scientific knowledge [35]. As future work, there is a need to create trustful models that allow the clinicians to understand which features contribute to the malignancy prediction [35]. However, this work showed important results to prove that this approach can have several advantages compared to other machine learning and deep learning solutions, maintaining the performance level for lung nodule classification.

5. Conclusions

We developed and applied an approach based on two steps: features extraction and classification, to help the diagnosis of lung nodules in CT images. This work was motivated by the need to explore options to overcome the lack of annotated biomedical data, which have been limiting the development of robust AI-based solutions in the medical field. In conclusion, this work showed that feature learning achieved when reconstructing the input with an encoder-decoder based architecture can be considered as useful knowledge in a transfer learning approach. This approach allows the use of data to learn without labeling constraints, which is one of the biggest limitations when using medical data, since the annotation is an expensive and extremely complex process.

Author Contributions: F.S., T.P., A.C. and H.P.O. conceived the scientific idea, C.F. and V.H. gave the pneumology insights about the malignancy risk assessment, and J.L.C. gave the molecular biology insights. F.S. conducted all the experiments. F.S., T.P., J.F., J.M., A.C. and H.P.O. contributed to the critical analysis of the results. F.S. and T.P. drafted the manuscript. All authors provided critical feedback and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financed by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020.

Acknowledgments: We acknowledged the National Cancer Institute and the Foundation for the National Institutes of Health for the free publicly available LIDC-IDRI Database used in this work.

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

References

1. World Health Organisation. *Latest Global Cancer Data: Cancer Burden Rises to 18.1 Million New Cases and 9.6 Million Cancer Deaths in 2018*; International Agency for Research on Cancer: Lyon, France, 2018.
2. American Cancer Society. *Facts & Figures 2019*; Technical Report; American Cancer Society: Atlanta, GA, USA, 2019.
3. Riquelme, D.; Akhloufi, M.A. Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans. *AI* **2020**, *1*, 28–67. [CrossRef]
4. Armato, S.G., III; McLennan, G.; Bidaut, L; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. Data From LIDC-IDRI. The Cancer Imaging Archive: Little Rock, AR, USA, 2015. [CrossRef]
5. Armato, S.G., III; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef] [PubMed]
6. Shen, W.; Zhou, M.; Yang, F.; Yang, C.; Tian, J. *Multi-Scale Convolutional Neural Networks for Lung Nodule Classification*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2015. [CrossRef]
7. Liu, L.; Liu, Y.; Zhao, H. Benign and malignant solitary pulmonary nodules classification based on CNN and SVM. In Proceedings of the ACM International Conference Proceeding Series, Singapore, 23–25 April 2018. [CrossRef]
8. Yan, X.; Pang, J.; Qi, H.; Zhu, Y.; Bai, C.; Geng, X.; Liu, M.; Terzopoulos, D.; Ding, X. Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 91–101. [CrossRef]
9. Song, Q.Z.; Zhao, L.; Luo, X.K.; Dou, X.C. Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images. *J. Healthc. Eng.* **2017**, *2017*, 8314740. [CrossRef] [PubMed]
10. 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*, 102–110. [CrossRef]
11. Causey, J.L.; Zhang, J.; Ma, S.; Jiang, B.; Qualls, J.A.; Politte, D.G.; Prior, F.; Zhang, S.; Huang, X. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci. Rep.* **2018**, *8*, 9286. [CrossRef] [PubMed]
12. Rizzo, S.; Botta, F.; Raimondi, S.; Origgi, D.; Fanciullo, C.; Morganti, A.G.; Bellomi, M. Radiomics: The facts and the challenges of image analysis. *Eur. Radiol. Exp.* **2018**, *2*, 36. [CrossRef] [PubMed]
13. Soleymani, S.; Dabouei, A.; Kazemi, H.; Dawson, J.; Nasrabadi, N.M. Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification. In Proceedings of the International Conference on Pattern Recognition, Beijing, China, 20–24 August 2018. [CrossRef]
14. Lindsay, W.; Wang, J.; Sachs, N.; Barbosa, E.; Gee, J. *Transfer Learning Approach to Predict Biopsy-Confirmed Malignancy of Lung Nodules from Imaging Data: A Pilot Study*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2018. [CrossRef]
15. Da Nóbrega, R.V.M.; Peixoto, S.A.; Da Silva, S.P.P.; Filho, P.P.R. Lung Nodule Classification via Deep Transfer Learning in CT Lung Images. In Proceedings of the International Symposium on Computer-Based Medical Systems (CBMS), Karlstad, Sweden, 18–21 June 2018; pp. 244–249. [CrossRef]
16. da Nóbrega, R.V.M.; Rebouças Filho, P.P.; Rodrigues, M.B.; da Silva, S.P.; Dourado Júnior, C.M.; de Albuquerque, V.H.C. Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks. *Neural Comput. Appl.* **2020**, *32*, 11065–11082. [CrossRef]
17. ImageNet. Available online: http://www.image-net.org/ (accessed on 27 January 2020).
18. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]

19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Zhang, Y.; Zhang, J.; Zhao, L.; Wei, X.; Zhang, Q. Classification of Benign and Malignant Pulmonary Nodules Based on Deep Learning. In Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE), Zhengzhou, China, 20–22 July 2018. [CrossRef]
21. Shi, Z.; Hao, H.; Zhao, M.; Feng, Y.; He, L.; Wang, Y.; Suzuki, K. A deep CNN based transfer learning method for false positive reduction. *Multimed. Tools Appl.* **2019**, *78*, 1017–1033. [CrossRef]
22. Cavallari, G.; Ribeiro, L.; Ponti, M. Unsupervised Representation Learning Using Convolutional and Stacked Auto-Encoders: A Domain and Cross-Domain Feature Space Analysis. In Proceedings of the 31st Conference on Graphics, Patterns and Images, (SIBGRAPI), Parana, Brazil, 29 October–1 November 2018. [CrossRef]
23. Kumar, D.; Wong, A.; Clausi, D.A. Lung Nodule Classification Using Deep Features in CT Images. In Proceedings of the 12th Conference on Computer and Robot Vision, Halifax, NS, Canada, 3–5 June 2015; pp. 133–138. [CrossRef]
24. Cheng, J.Z.; Ni, D.; Chou, Y.H.; Qin, J.; Tiu, C.M.; Chang, Y.C.; Huang, C.S.; Shen, D.; Chen, C.M. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Sci. Rep.* **2016**, *6*, 24454. [CrossRef] [PubMed]
25. Wiemker, R.; Bergtholdt, M.; Dharaiya, E.; Kabus, S.; Lee, M.C. Agreement of CAD features with expert observer ratings for characterization of pulmonary nodules in CT using the LIDC-IDRI database. *SPIE Med. Imaging* **2009**, *7260*, 72600. [CrossRef]
26. Lin, H.; Huang, C.; Wang, W.; Luo, J.; Yang, X.; Liu, Y. Measuring Interobserver Disagreement in Rating Diagnostic Characteristics of Pulmonary Nodule Using the Lung Imaging Database Consortium and Image Database Resource Initiative. *Acad. Radiol.* **2017**, *24*, 401–410. [CrossRef] [PubMed]
27. Nibali, A.; He, Z.; Wollersheim, D. Pulmonary nodule classification with deep residual networks. *Int. J. Comput. Assist. Radiol. Surg.* **2017**, *12*, 1799–1808. [CrossRef] [PubMed]
28. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6. [CrossRef]
29. Wang, Y.; Yao, H.; Zhao, S.; Zheng, Y. Dimensionality reduction strategy based on auto-encoder. In Proceedings of the ACM International Conference Proceeding Series, Zhangjiajie, China, 19–21 August 2015. [CrossRef]
30. Pihlgren, G.G.; Sandin, F.; Liwicki, M. Improving Image Autoencoder Embeddings with Perceptual Loss. *arXiv* **2020**, arXiv:cs.CV/2001.03444.
31. Alain, G.; Bengio, Y. What Regularized Auto-Encoders Learn from the Data Generating Distribution. *arXiv* **2012**, arXiv:cs.LG/1211.4246.
32. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss Functions for Image Restoration with Neural Networks. *IEEE Trans. Comput. Imaging* **2016**, *3*, 47–57. [CrossRef]
33. 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*, 226–238. [CrossRef] [PubMed]
34. Ichimura, N. Spatial Frequency Loss for Learning Convolutional Autoencoders. *arXiv* **2018**, arXiv:cs.CV/1806.02336.
35. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]