

Lung Nodule Classification for Clinical Decision Support Systems: A Comparative Analysis of DenseNet-121 and Xception Architectures

Mesut ULU ^{1*} 

¹ Bandırma Onyedi Eylül University, Manyas Vocational School, Department of Occupational Health and Safety, Balıkesir, Türkiye

Abstract

The early and accurate assessment of lung nodules is an important clinical requirement worldwide, as it contributes to the timely evaluation of potential malignancy risk. In this study, DenseNet-121 and Xception architectures were compared for the binary classification of individual two-dimensional (2D) computed tomography (CT) slices into Healthy and Lung_Nodule categories, rather than for patient-level diagnosis or malignancy risk assessment. A total of 20,432 images were evaluated in two classes, Healthy and Lung_Nodule, and both models were trained using a transfer learning approach under the same data-splitting and preprocessing conditions. DenseNet-121 and Xception exhibited broadly parallel overall performance, with DenseNet-121 achieving an accuracy of 0.894, a balanced accuracy of 0.894, and an F1-score of 0.894. For Xception, these values were obtained as 0.892, 0.892, and 0.892, respectively. When the clinical error profiles were examined, DenseNet-121 provided higher Specificity (0.946) and a lower FPR (0.054), whereas Xception showed a small advantage in reducing missed nodule images with a lower FNR (0.152). Grad-CAM visualizations provided qualitative insights into the image regions emphasized by the models in selected correct and incorrect predictions. The findings suggest that, despite their similar overall performance, the two architectures differ in terms of their clinical error behavior.

Keywords: Lung nodule; Deep learning; Transfer learning; Computed tomography; Decision support system.

1. Introduction

Lung nodules are among the pulmonary findings that may be encountered in computed tomography (CT) examinations and require careful clinical evaluation. Although the presence of a nodule alone does not necessarily indicate malignancy, assessing the potential risk at an early stage is an important part of patient management. In fact, the National Lung Screening Trial demonstrated that lung cancer screening with low-dose CT can reduce mortality in high-risk individuals, making the role of CT imaging in the early assessment process more evident [1]. Since the size, number, and imaging characteristics of pulmonary nodules may influence the follow-up approach, systematic evaluation of nodules detected on CT is important for clinical decision-making [2].

The evaluation of CT examinations is not always straightforward. Nodules can exhibit varying appearances in terms of size, shape, margin, and density, and they may also produce imaging characteristics similar to those of vascular and other anatomical structures. Even among experienced thoracic radiologists identifying nodules on the same CT images, interobserver variability has been reported to arise [3]. This situation makes the use of computer-aided systems as a second evaluation tool particularly meaningful in clinical settings where large numbers of images are assessed. The primary aim here is not to replace the clinician but rather to provide a decision support that can assist in the more consistent identification of suspicious patterns within the image.

Deep learning at this point has brought a significant methodological advantage to medical image analysis. Whereas in traditional image analysis features largely need to be predefined, convolutional neural networks are able to learn low- and high-level features from images during the training process. CNN-based approaches have been widely used across different medical imaging problems such as classification, detection, and segmentation, and have become a powerful approach particularly in terms of automatic feature extraction [4]. However high classification performance alone does not indicate that a model is suitable for clinical use. In medical artificial intelligence research, limited external validation, performance variations across different datasets, and the heterogeneous nature of comparisons conducted with clinicians represent important challenges [5-6].

A similar level of variation is also seen in deep learning studies focusing on pulmonary nodules. Differences in the datasets used, classification targets, image sizes, preprocessing procedures, and model configurations may make direct comparison between architectures difficult. The systematic review by Li et al. (2019) [6] particularly highlighted these performance and validation issues in pulmonary nodule models tested across

*Corresponding author

E-mail address: mulu@bandirma.edu.tr

different datasets. In addition, although overall metrics such as Accuracy or AUC are emphasized in many studies, interpreting false-positive and false-negative outcomes with the same weight may obscure the actual error behavior of a model, since these errors have different clinical implications. In particular, classifying an image containing a nodule as healthy represents an error type that deserves separate attention in the context of a clinical decision support system.

Within this framework, the main aim of the study is to compare the classification performances of DenseNet-121 and Xception architectures on CT images consisting of Healthy and Lung_Nodule classes under the same experimental conditions, and to examine not only their overall performance but also their clinically meaningful error patterns. Both models were evaluated using the same data-splitting, preprocessing, and training strategy. In addition to basic metrics such as Accuracy, Precision, Recall, F1-Score, and Specificity, NPV, FPR, FNR, and confusion matrix results were also considered. Furthermore, Grad-CAM was used to examine which regions of the images attracted the models' attention in both correct and incorrect predictions. Accordingly, this study seeks to answer three research questions: RQ1: Do the DenseNet-121 and Xception architectures differ in terms of overall classification performance under the same data and training conditions? RQ2: How do the models differ in their sensitivity, specificity, and particularly FP–FN error patterns for the Healthy and Lung_Nodule classes? RQ3: Can Grad-CAM visualizations provide explanatory support regarding the regions on which the models focus in correct and incorrect classifications?

The contribution of this study can be summarized in four points. First, two structurally different transfer learning architectures were compared within a common experimental framework, allowing architecture-dependent performance differences to be observed more clearly. Second, the evaluation was not limited to Accuracy, but was extended with class-based and clinically relevant diagnostic metrics. Third, false-positive and particularly false-negative outcomes were examined separately to reveal the clinical error profiles of the models. Finally, Grad-CAM analysis was used to support the numerical performance results from the perspective of visual interpretability.

2. Literature Review

Computed tomography (CT) is among the commonly employed imaging modalities for the detection and assessment of pulmonary nodules. However the examination of numerous cross-sectional slices, the possibility that small nodules may be overlooked, and the fact that certain anatomical structures can mimic nodular appearances, all have the potential to complicate the evaluation process. For this reason computer-aided diagnosis and clinical decision-support systems have gained growing importance as complementary tools that assist radiological assessment. Masood and colleagues (2020) [7], having developed a 3D CNN-based decision-support system aimed at pulmonary nodule detection and diagnosis in chest CT images, demonstrated that deep learning holds practical utility within the clinical evaluation setting.

CNN-based approaches have been used at different stages of automatic lung nodule analysis. Setio et al. (2016) [8] employed a multi-view CNN structure to reduce false positives and showed that information obtained from different imaging planes contributed to distinguishing true nodules from false-positive candidates. Dou et al. (2017) [9] evaluated the spatial information surrounding nodule candidates using a multi-level contextual 3D CNN architecture, with particular emphasis on reducing false positives. Nasrullah et al. (2019) also proposed a multi-stage deep learning approach that addressed nodule detection and classification together. These studies indicate that not only the appearance of the nodule itself, but also the contextual structures around it, may play an important role in the classification process [8–10].

In nodule classification, the concurrent evaluation of different scales and appearances is another approach that has drawn attention in the literature. Shen and colleagues (2017) [11], using a multi-crop CNN architecture, assessed nodule features across multiple scales and examined the classification of malignancy suspicion. Xie and colleagues (2019) [12] on the other hand developed a knowledge-based collaborative deep learning model that handles the 3D nodule structure from nine fixed viewpoints. Similarly, Mastouri et al. (2021) [13] modeled different feature interactions together using a bilinear CNN and presented an alternative approach for nodule classification in CT images. These studies suggest that moving beyond information obtained from a single image level may be useful for nodule characterization.

The utilization of the volumetric nature of CT data also represents an important research direction in pulmonary nodule studies. DeepLung, developed by Zhu and colleagues (2018) [14], was presented as an integrated system that combines nodule detection and classification within 3D deep networks. Ardila and colleagues (2019) [15] on the other hand developed a three-dimensional deep learning model that evaluates low-dose CT images in an end-to-end manner and achieved an AUC of 0.944 on 6,716 NLST cases. Venkadesh and colleagues (2021) [16] also made use of both 2D and 3D image information to predict the malignancy risk of pulmonary nodules detected on screening CT, reporting that model performance was comparable to that of thoracic radiologists in independent validation. These findings highlight the importance of modeling particularly the three-dimensional morphology of the nodule and its relationship with surrounding tissue.

In recent years the performance of certain CNN architectures in pulmonary nodule analysis has also been

investigated separately. Canayaz and colleagues (2024) [17] addressed pulmonary nodule classification and segmentation together in CT images and evaluated DenseNet-121 as one of the backbone models employed. In their study DenseNet-121 was utilized particularly during the feature extraction stage. Li and colleagues (2022) [18], who examined the use of the Xception architecture in nodule classification, proposed the DCA-Xception model which combines dual-channel feature extraction and an attention mechanism with the Xception structure. In their study where benign and malignant nodules were distinguished, an AUC value of 0.929 was reported. These findings suggest that DenseNet- and Xception-based structures offer two distinct architectural approaches that can be used for learning features from pulmonary CT images.

When the literature is considered as a whole, a substantial part of the existing studies appears to focus on nodule detection, false-positive reduction, or benign–malignant differentiation. In addition to multi-view and three-dimensional networks, the use of different feature extraction mechanisms has also become increasingly common. In the present study, however, the problem was addressed at a more basic level of clinical pre-assessment, with the aim of distinguishing Healthy and Lung_Nodule images. DenseNet-121 and Xception architectures were compared under the same data-splitting and preprocessing conditions, allowing their performance and error behaviors to be evaluated directly. Particular attention was also given to false-negative results, making the risk of classifying images containing nodules as healthy an explicit part of the model comparison.

3. Materials and Methods

3.1. Workflow of the study

The overall methodological structure of the study has been planned in successive stages, ranging from the data preparation process through model training and clinical performance evaluation. In order to enable a comparison of the DenseNet-121 and Xception architectures under identical experimental conditions, the entire workflow was conducted over common data splitting, preprocessing, and evaluation steps. Figure 1 illustrates the experimental process followed in the study holistically, from the data source through to clinical assessment. In the first stage the dataset consisting of 20,432 CT images was partitioned into training, validation, and test sets, after which the images were resized to 224×224 pixels and subjected to RGB conversion, normalization, and data augmentation procedures. The prepared images were then fed separately into the DenseNet-121 and Xception architectures using an ImageNet-based transfer learning approach. Evaluating both models under the same data structure and training conditions allows for a more direct comparison of performance differences between the architectures.

After model training, the images were classified into Healthy and Lung_Nodule classes, and performance was evaluated using Accuracy, Recall, Specificity, F1-Score, and FP/FN values. In addition, confusion matrices were used to examine the error types numerically, while Grad-CAM analysis was applied to visually assess which image regions the models focused on during their classification decisions. In this way, the workflow provides a comparative evaluation that considers not only overall classification performance, but also the error behavior of the models and their tendency to miss nodule-containing images, which is clinically important.

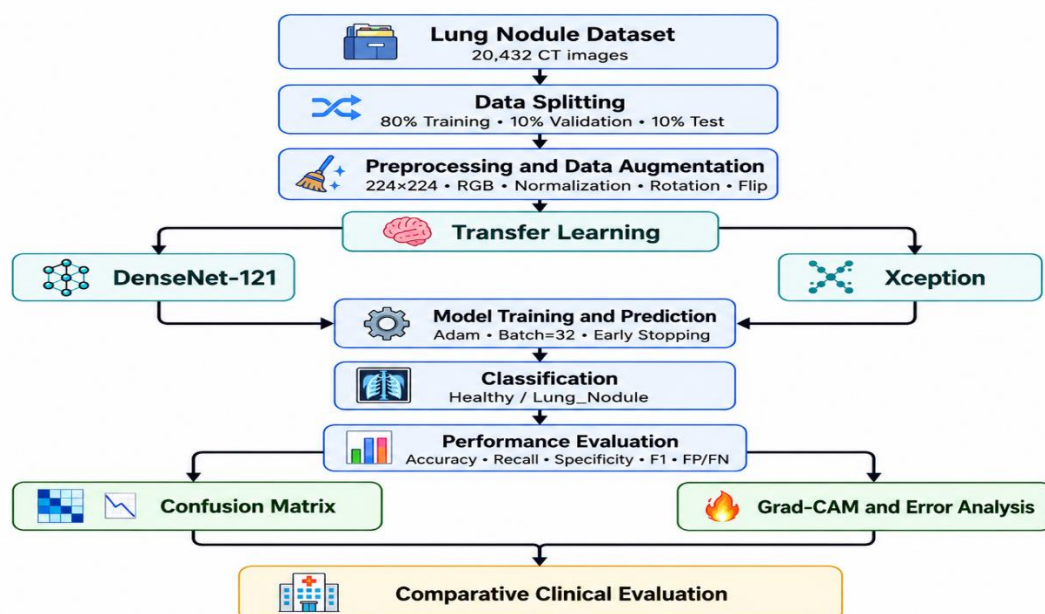


Figure 1. Overall workflow of the study

3.2. Data source and data collection process

In this study the Lung Nodule Dataset, publicly available on the Kaggle platform was used for the purpose of classifying pulmonary nodules from computed tomography (CT) images [19]. The dataset was accessed through the relevant repository on Kaggle, and no new image collection process from patients was conducted directly within the scope of the study. For this reason the data utilized consist of publicly available secondary medical imaging data. The dataset contains a total of 20,432 CT images organized into two classes: Healthy and Lung_Nodule. The Healthy class refers to CT slices labeled as not containing visible pulmonary nodules (nodule-free slices) and does not necessarily indicate that the corresponding individuals are clinically healthy. In contrast, the Lung_Nodule class represents CT slices labeled as containing pulmonary nodules. The original class names were retained to maintain consistency with the dataset labels. It is important to emphasize that the unit of analysis was an individual two-dimensional (2D) CT slice rather than a complete CT examination or an individual patient. Therefore, the classification results do not represent patient-level diagnosis, precise nodule localization, or malignancy risk assessment. The distribution of the two classes within the dataset being fairly balanced also reduces the risk of a pronounced class dominance arising during model training. The dataset comprises a total of 20,432 CT images, including 10,292 Healthy images and 10,140 Lung_Nodule images. The Healthy class represents CT slices labeled as not containing visible pulmonary nodules, whereas the Lung_Nodule class represents slices labeled as containing pulmonary nodules. However, detailed information regarding the original clinical institution, the number of unique patients, and the initial annotation and verification procedures was not available in the dataset documentation. Furthermore, radiologist-validated nodule annotations and histopathological confirmation could not be verified. Therefore, the class labels should be interpreted as dataset-provided image-level categories rather than independently confirmed clinical diagnoses.

The general structure of the images used in the study is shown in Figure 2. In the figure, CT images representing the Healthy and Lung_Nodule classes are presented together, with the aim of providing an overall view of the visual distinction between the two classes. Particularly in slices containing nodules, the local density differences observed within the lung parenchyma constitute one of the primary image features that the classification models attempt to learn.

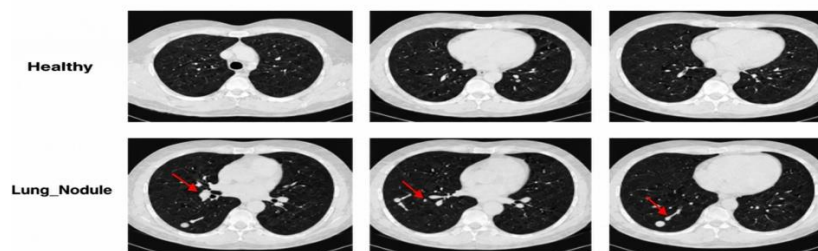


Figure 2. Showing example CT images belonging to the Healthy (nodule-free) and Lung_Nodule classes

The entire dataset was split into training, validation, and test sets prior to modeling. To ensure that the splitting process yielded the same result across different runs and to preserve the reproducibility of the experiments, the randomness value was set to seed=42. Eighty percent of the dataset was used for training, ten percent for validation, and ten percent for testing. Following this procedure 16,345 images were allocated to the training set, 2,043 to the validation set, and 2,044 to the test set. The presence of 1,030 Healthy and 1,014 Lung_Nodule images within the test set also ensured that the classes exhibited a closely balanced distribution during the evaluation phase. The dataset was divided at the individual image (2D CT slice) level using a fixed random seed of 42. Since the publicly available dataset did not provide patient-level or examination-level identifiers for grouping, patient-wise data splitting could not be performed. Accordingly, the reported results represent image-level classification performance rather than performance evaluated on independent patients.

3.3. Data Preprocessing

In order to enable a comparison of the deep learning models under identical conditions, a common preprocessing pipeline was applied to the images fed into the DenseNet-121 and Xception architectures. In the first stage the raw CT images were resized to 224×224 pixels so as to be compatible with the input layers of the models. Although the images essentially carry grayscale information, they were fed into the models in RGB format due to the three-channel input structure of the DenseNet-121 and Xception architectures, both of which had been pretrained on the ImageNet dataset. To bring the pixel values into a common range, the intensity values between 0 and 255 were rescaled using a factor of $1/255$ and normalized to the 0–1 range. A common normalization strategy based on a scaling factor of $1/255$ was adopted for both DenseNet-121 and Xception to maintain consistent input preprocessing conditions and facilitate a controlled comparison between the

architectures. Accordingly, pixel values were scaled to the range of 0–1, without applying architecture-specific ImageNet preprocessing functions. Although DenseNet-121 typically employs ImageNet-based channel normalization and Xception uses input scaling to the range of -1 to 1, the common normalization approach was preferred to maintain an identical preprocessing pipeline for both models. However, this choice may have influenced the feature extraction capabilities of the pretrained architectures and is therefore acknowledged as a methodological limitation of the study.

Limited data augmentation procedures were also applied to the training set in order to reduce the likelihood of the model learning features dependent solely on the position or orientation of the images. Within this framework random rotation up to 10 degrees (`rotation_range=10`) and horizontal flipping (`horizontal_flip=True`) were applied to the images. Data augmentation was used only on the training images, whereas the original structure of the images was preserved for the validation and test sets. In this way while the models encountered a wider range of image variations during training, performance evaluation was carried out on images that had not undergone any geometric transformation.

The images were provided to the model in mini-batches of 32 samples. In the test data generator, the `shuffle=False` option was used to preserve the order between the true class labels and the prediction results. This detail is particularly important for the correct calculation of the confusion matrix, true positive, false positive, and false negative values. By applying the same preprocessing steps to both models, DenseNet-121 and Xception were compared under more consistent experimental conditions.

3.4. DenseNet-121 architecture

DenseNet-121 is a 121-layer architecture belonging to the DenseNet family, which was developed to render information flow between layers more efficient in deep convolutional neural networks. Whereas in conventional CNN structures a layer typically utilizes the output of the immediately preceding layer, within the DenseNet architecture each layer can directly access the feature maps generated by all previous layers. Through this dense connectivity structure, features learned in the earlier layers of the network are carried forward to subsequent layers, thereby enabling feature reuse. At the same time this structure contributes to mitigating the vanishing gradient problem that can be encountered in deep networks and to preserving information flow throughout the network [20].

In the DenseNet-121 architecture, the layers are organized within densely connected structures known as dense blocks. As the feature maps obtained from one layer are transferred to subsequent layers, the previously learned features are not completely lost and continue to be used together with newly extracted information. This is particularly important for preserving image characteristics learned at different levels, such as edges, textures, and shapes. The transition layers between dense blocks help reduce the dimensions of the feature maps and thereby control the computational load of the network [20].

The applicability of this architectural structure to medical imaging has also been demonstrated in different studies. Particularly in the evaluation of nodules on lung CT images, not only prominent shape characteristics but also texture and intensity differences surrounding the nodule are important. In studies analyzing pulmonary nodules from CT images, DenseNet-121 has been considered one of the main models used for feature extraction. The ability of DenseNet-121 to transfer features learned in earlier layers to subsequent layers may provide an advantage in jointly evaluating visual information available at different levels in medical images [17].

In this study, DenseNet-121 was selected for the classification of CT images containing lung nodules and healthy images. Its densely connected structure was considered potentially useful for learning subtle tissue differences in CT images as well as morphological characteristics around the nodules. In addition, by using weights pretrained on ImageNet, the model benefited from previously learned general image features and was adapted to the current dataset through a transfer learning approach.

3.5. Xception architecture

Xception is a deep convolutional neural network architecture developed by Chollet (2017) [21] that approaches the Inception paradigm from a different perspective. The architecture largely employs depthwise separable convolution operations in place of classical convolution operations. In this approach the learning of spatial features and the learning of inter-channel relationships are disentangled from one another. Spatial convolution is first applied separately over each channel, following which the information across channels is aggregated through 1×1 pointwise convolution. In this way a structure is obtained where parameters are utilized more efficiently compared to standard convolution, while strong representational capacity is nonetheless preserved [21].

The basic idea behind the Xception architecture is based on the assumption that spatial correlations and cross-channel correlations in images can be modeled separately. For this reason, the architecture may be considered a more advanced and simplified form of Inception modules. In addition, the residual connections used within the network help information and gradients to be transferred more steadily through deeper layers.

In the original Xception study, the architecture was also reported to achieve competitive performance compared with Inception V3 of a similar parameter size, and even higher performance on some datasets in image classification tasks [21].

The usability of this structure for the classification of pulmonary nodules was also addressed in subsequent studies. With the DCA-Xception network they developed based on Xception, the classification of pulmonary nodules in CT images was examined. In their study dual-channel feature extraction and an attention mechanism were incorporated into the Xception structure, and an AUC value of 0.929 was reported with the proposed model. This result indicates that Xception-based structures can constitute a usable foundation in terms of discriminating image features in pulmonary nodules [18].

In the preference for the Xception architecture in this study, its capability to learn tissue and shape features present in CT images in a more disentangled manner was influential. Particularly since pulmonary nodules can manifest in small regions, it is important not only to learn the overall image structure but also local tissue and density variations. It was thought that the depthwise separable convolution structure of Xception might be suitable for extracting such features, and the model was compared with DenseNet-121 under the same dataset and identical experimental conditions.

3.6. Evaluation metrics

The performances of the DenseNet-121 and Xception models were evaluated using Accuracy, Precision, Recall, Specificity, and F1-Score metrics. In addition the number of False Negative (FN) cases, which indicates images containing pulmonary nodules being classified as healthy, was also taken into consideration separately. Since overall accuracy alone is not sufficient in medical image classification, the models were examined together across different performance criteria.

In this study, the Lung_Nodule class was considered as the positive class. Recall indicates how many of the actual nodule-containing images were correctly identified, while Precision reflects the correctness of the images predicted as Lung_Nodule by the model. Specificity represents the proportion of healthy images that were correctly classified. F1-Score evaluates Precision and Recall together, whereas Accuracy gives the proportion of total correct predictions among all samples. A decision threshold of 0.50 was used to convert the model outputs into class labels. Accordingly, predictions above 0.50 were classified as Lung_Nodule, while lower values were assigned to the Healthy class.

The performance metrics were calculated using the following equations[22]:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$F1 - Score = 2 * \frac{(Precision*Recall)}{(Precision+Recall)} \quad (4)$$

$$Specificity = \frac{TN}{(FP+TN)} \quad (5)$$

Through the joint evaluation of these metrics, not only the overall success of the models but also their classification behavior on nodule-bearing and healthy images was compared.

4. Experimental Results

The experimental performances of DenseNet-121 and Xception were compared under the same data-splitting structure and preprocessing conditions. Both models were initialized with weights pretrained on ImageNet, while the convolutional base layers were frozen to implement a transfer learning approach. During training, the Adam optimization algorithm was used with a learning rate of 0.001, and the mini-batch size was set to 32. To reduce the risk of overfitting, validation loss was monitored, and an EarlyStopping mechanism was employed to terminate the training process when no improvement was observed for three consecutive epochs. The experiments were conducted in Python using Jupyter Notebook, with TensorFlow 2.15.0 and Keras 2.15.0. Both DenseNet-121 and Xception were configured for binary classification of CT images into Healthy and Lung_Nodule classes. The binary cross-entropy loss function was employed, and the maximum number of training epochs was set to 50. The same training configuration, including the optimizer, learning rate, batch size, loss function, and EarlyStopping settings, was applied to both architectures to ensure a consistent comparison. For both architectures, the classification head consisted of a Global Average Pooling layer, a fully connected Dense layer with 256 units and ReLU activation, a Dropout layer with a rate of 0.5, and a single-unit sigmoid output layer. ModelCheckpoint (save_best_only=True) was used to save the model weights corresponding to the lowest validation loss, and the best checkpoint was used for test evaluation. The experiments were conducted on an NVIDIA RTX 3090 GPU (24 GB VRAM), an Intel Core i9-12900K CPU, and 64 GB RAM, using Ubuntu 22.04 and Python 3.10. No additional fine-tuning was performed, and only the newly added classification head was trained.

The performances of the models on the test set were evaluated using Accuracy, Precision, Recall, Specificity, and F1-Score. In addition, false-negative predictions, which are considered more critical from a clinical perspective, were examined separately. Evaluating both architectures on the same test set was preferred to ensure that the observed performance differences were more closely related to the model architecture rather than to variations in the data or experimental conditions. The overall classification results of DenseNet-121 and Xception are presented comparatively in Table 1.

Table 1. Overall Classification Performance of the Models

Performance Metric	DenseNet-121	Xception
Accuracy	0.894	0.892
Balanced Accuracy	0.894	0.892
Precision (macro)	0.899	0.895
Recall (macro)	0.894	0.892
F1-Score (macro)	0.894	0.892

When Table 1 is examined, it is seen that the DenseNet-121 and Xception architectures exhibited performances quite close to one another in pulmonary nodule classification. DenseNet-121 achieved an Accuracy of 0.894, compared with 0.892 for Xception, corresponding to an absolute difference of approximately 0.2 percentage points. This difference is descriptive and very limited in magnitude. Since confidence intervals and paired statistical significance tests were not performed, the observed difference should not be interpreted as evidence of statistically significant superiority of either architecture. A similar tendency was also observed in the Balanced Accuracy results. The Balanced Accuracy value of 0.894 for DenseNet-121 was calculated as 0.892 for Xception. The closeness of the Accuracy and Balanced Accuracy values indicates that the models' performance across the two classes was generally well-balanced. Particularly in medical image analyses where misclassifications such as those involving pulmonary nodules can carry significant consequences, it is important not only to consider overall accuracy but also to maintain this balance between classes.

In terms of Precision, DenseNet-121 achieved 0.899, which was slightly higher than the 0.895 obtained by Xception. Recall values were 0.894 and 0.892, respectively. The F1-Score values of 0.894 for DenseNet-121 and 0.892 for Xception also indicate that both architectures maintained a similar balance between Precision and Recall. Overall, although DenseNet-121 produced slightly higher performance values, the difference between the two architectures remained quite small. Therefore, to better reveal clinically important differences in lung nodule classification, class-based Recall, Specificity, and false-negative results should also be considered together with the general performance metrics.

Table 2 reveals that the class-based performances of the DenseNet-121 and Xception models were generally quite close to one another, yet the two architectures exhibited differing tendencies with respect to false positive and false negative errors. When the results for the Healthy class are examined, the Healthy Precision value is seen to be 0.859 for DenseNet-121 and 0.862 for Xception. In this regard Xception demonstrates a very marginal superiority. On the other hand the Healthy Recall value was obtained as 0.946 for DenseNet-121 whereas it was 0.936 for Xception. This result indicates that DenseNet-121 holds an advantage of approximately one percentage point in correctly identifying genuine healthy images. In other words DenseNet-121 keeps the tendency to mistakenly classify healthy images as nodule-bearing at a lower level compared to Xception.

Table 2. Class-Based Performance of the Models

Performance Metric	DenseNet-121	Xception
Healthy Precision	0.859	0.862
Lung_Nodule Precision	0.938	0.929
Healthy Recall	0.946	0.936
Lung_Nodule Recall	0.842	0.848
Healthy F1-Score	0.900	0.898
Lung_Nodule F1-Score	0.888	0.887

A different performance balance emerges for the Lung_Nodule class. The Lung_Nodule Precision value was calculated as 0.938 for DenseNet-121 and 0.929 for Xception. Accordingly, images classified as nodules by DenseNet-121 were more likely to actually contain a nodule. This indicates that DenseNet-121 showed a more selective classification behavior in limiting false-positive predictions. In contrast, the Lung_Nodule

Recall value was 0.842 for DenseNet-121 and 0.848 for Xception, suggesting that Xception had a limited advantage of approximately 0.6 percentage points in detecting images with actual nodules. From the perspective of a clinical decision support system, this difference is important because higher nodule sensitivity means a lower risk of incorrectly classifying nodule-containing images as healthy.

The F1-Score results on the other hand indicate that the difference between the two models is quite limited. For the Healthy class DenseNet-121 and Xception achieved F1-Score values of 0.900 and 0.898 respectively, while for the Lung_Nodule class these values were determined as 0.888 and 0.887 respectively. Thus in terms of F1-Score, which considers both precision and recall together, there is no pronounced performance divergence between the two architectures. Although DenseNet-121 shows a very small superiority in both classes, the magnitude of the differences is practically limited.

Table 3 presents the image-level classification metrics of the two models, highlighting small differences in their false-positive and false-negative error patterns rather than demonstrating clinical diagnostic performance. DenseNet-121, with a Specificity value of 0.946, is successful than Xception in correctly distinguishing healthy images. Consistent with this, the False Positive Rate being 0.054 for DenseNet-121 and 0.064 for Xception indicates that DenseNet-121 has a slightly lower tendency to mistakenly evaluate a healthy image as nodule-bearing. This can be considered a result in favor of DenseNet-121 in terms of reducing unnecessary positive alerts.

In contrast the Xception model showed a small advantage in terms of not overlooking nodule-bearing images. The False Negative Rate was calculated as 0.158 for DenseNet-121 and 0.152 for Xception. Accordingly Xception kept the rate of classifying genuine nodule cases as healthy slightly lower. Similarly the NPV value being 0.862 for Xception and 0.859 for DenseNet-121 indicates that the model's predictions labeled as healthy are marginally more reliable on the Xception side. From the perspective of clinical decision-support systems this distinction is important. While DenseNet-121 appears stronger in reducing false positives and distinguishing healthy cases, Xception provided a slight advantage in terms of reducing false negatives—that is, in not missing nodule-containing images.

Table 3. Image-Level Classification Error Metrics of the Models

Performance Metric	DenseNet-121	Xception
Specificity	0.946	0.936
Negative Predictive Value (NPV)	0.859	0.862
False Positive Rate (FPR)	0.054	0.064
False Negative Rate (FNR)	0.158	0.152

The confusion matrix for DenseNet-121 presented in Figure 3 shows that the model correctly classified 974 of the 1,030 healthy images in the test set, while erroneously labeling 56 healthy images as nodule-bearing. On the other hand 854 of the 1,014 nodule-containing images were correctly detected, while 160 nodule-bearing images were predicted as healthy. This distribution indicates that the model performed strongly in distinguishing the Healthy class and achieved a high level of specificity, although some images containing nodules were missed. At the image level, DenseNet-121 produced 56 false-positive and 160 false-negative predictions, corresponding to a false-negative rate of 15.8%. This finding indicates that a considerable proportion of nodule-containing CT slices were incorrectly classified as Healthy. Therefore, the observed performance should not be interpreted as demonstrating sufficient reliability for clinical application, particularly in the absence of patient-level and external clinical validation.

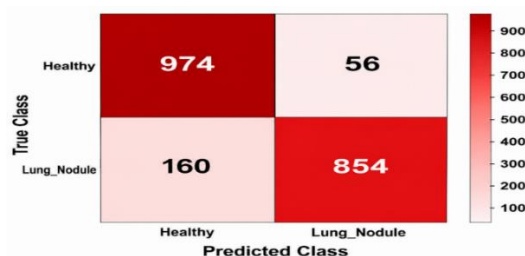


Figure 3. Confusion matrix of the DenseNet-121 model

The confusion matrix for the Xception model given in Figure 4 shows that the model correctly classified 964 of the 1,030 healthy images in the test set, while erroneously labeling 66 healthy images as nodule-bearing. For the nodule-containing class, 860 of the 1,014 images were correctly detected, while 154 images were predicted as healthy. These results indicate that the Xception model exhibited a balanced performance across

both healthy and nodule classes, with a relatively strong true positive capture capacity particularly in the nodule class. The resulting false-negative rate of 15.2% indicates that a notable proportion of nodule-containing CT slices were incorrectly classified as Healthy. Although Xception produced six fewer false-negative predictions than DenseNet-121, this difference was relatively small and should be interpreted descriptively.

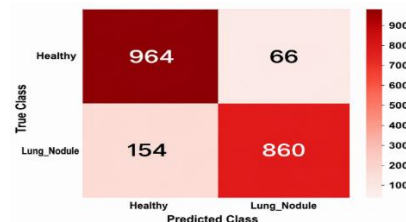


Figure 4. Confusion matrix of the Xception model

Figure 5 presents Grad-CAM visualizations for selected true-positive (TP) and false-negative (FN) predictions of DenseNet-121 and Xception. The examples were selected from the test set based on their classification outcomes to illustrate correctly identified and missed nodule-containing CT slices. These visualizations provide qualitative information about the image regions emphasized by the models rather than a quantitative assessment of nodule localization. Since the activation maps were not validated against expert-annotated nodule locations, they should not be interpreted as direct causal evidence of the models' decision-making processes. Numerically, Xception produced 860 true positives and 154 false negatives, compared with 854 true positives and 160 false negatives for DenseNet-121.

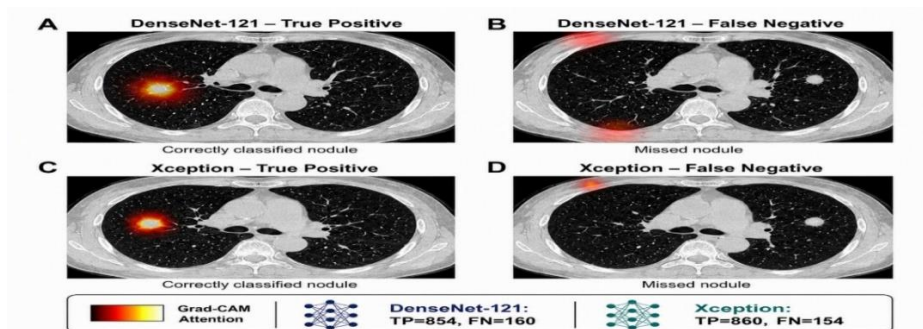


Figure 5. Comparison of Grad-CAM based models

Figure 6 presents a comparative display of the false positive (FP) and false negative (FN) error distributions for the DenseNet-121 and Xception models. For the DenseNet-121 model 56 FP and 160 FN were observed, whereas for the Xception model 66 FP and 154 FN were recorded. These results indicate that DenseNet-121 has a lower tendency to misclassify healthy images as nodule-bearing and is thus more successful in reducing false positives. Xception on the other hand provided a limited advantage in terms of not overlooking nodule-containing images, reducing the false negative count from 160 to 154. From a clinical standpoint false negative errors are considered more critical as they may lead to a genuine nodule being evaluated as healthy; therefore Xception's lower FN value is a noteworthy finding with respect to clinical decision support. In contrast DenseNet-121 producing fewer FPs offers an advantage in reducing unnecessary further examination or false alarm burden. Thus although the two architectures show similar overall performance, their error profiles diverge, with one exhibiting a slight superiority in terms of specificity and the other in terms of nodule sensitivity.

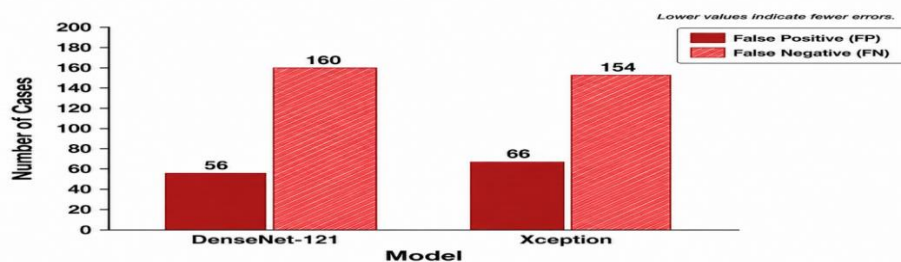


Figure 6. Error profiles of the models

The findings of the present study can also be considered in relation to previous deep learning approaches

for pulmonary nodule classification. Zhang et al. (2019) [23] reported Accuracy values of 87.77% for DenseNet-121 and 87.03% for Xception using three-dimensional CT nodule data. Similarly, Li et al. (2022) [18] obtained an Accuracy of 83.46% using a DCA-Xception architecture. In comparison, the present study achieved higher numerical Accuracy values of 89.4% and 89.2% for DenseNet-121 and Xception, respectively. Furthermore, the higher Specificity of DenseNet-121 and higher Sensitivity of Xception observed in the present study are consistent with the architectural performance tendencies reported by Zhang et al. However, differences in classification objectives, dataset characteristics, image dimensionality, and experimental protocols may account for variations in the reported results. Since the previous studies focused on benign–malignant nodule classification, whereas the present study addressed image-level Healthy–Lung_Nodule classification, these numerical comparisons should not be interpreted as evidence of direct methodological superiority.

5. Conclusion

In this study, the DenseNet-121 and Xception architectures were compared under identical data splitting, preprocessing, and training conditions for the classification of lung CT images as Healthy and Lung_Nodule. The results obtained showed that the overall performances of the two architectures were quite close. DenseNet-121 achieved Accuracy, Balanced Accuracy, and F1-Score values of 0.894, 0.894, and 0.894, respectively, whereas Xception achieved 0.892, 0.892, and 0.892. The observed performance differences were very small. Thus the results indicate that rather than a pronounced performance difference between the architectures, two different classification behaviors with similar overall success emerge.

When evaluated from a clinical standpoint the difference between the models was observed more in their error profiles. DenseNet-121 yielded a Specificity value of 0.946 and an FPR of 0.054, thereby demonstrating a lower tendency to falsely label healthy images as nodule-bearing. Xception on the other hand remained slightly below DenseNet-121's FNR of 0.158 with a value of 0.152, providing a small advantage in terms of not overlooking nodule-containing images. This tendency was also preserved in the confusion matrices; 160 false negative predictions were observed for DenseNet-121 and 154 for Xception. Grad-CAM visualizations provided qualitative information about the image regions emphasized by the models in selected correct and incorrect predictions. However, because these activation patterns were not compared with expert-annotated nodule locations, they are best interpreted as qualitative visual cues rather than definitive evidence of nodule localization.

The study has certain limitations. First and foremost the evaluation was conducted on a single publicly available dataset, and no external validation was performed on independent data obtained from a different clinical center. Since the problem addressed was limited to binary classification as Healthy–Lung_Nodule, the results should not be interpreted as benign–malignant differentiation of nodules or estimation of malignancy risk. Although the dataset consists of 20,432 CT images, the study was carried out on two-dimensional CT slices, and the three-dimensional structure of the nodule across consecutive sections was not directly incorporated into the model. Furthermore the fact that the models were evaluated with a fixed decision threshold of 0.50 indicates that different priorities between sensitivity and specificity in clinical use were not optimized within the scope of this study. The lack of patient-level identifiers precluded patient-wise data splitting; therefore, overlap across data subsets and potential information leakage cannot be ruled out. Future studies should use patient-level identifiers to enable independent evaluation and improve generalizability. In addition, limited information regarding the original clinical source, number of unique patients, and annotation procedures restricts the assessment of dataset provenance and label reliability. Overall, these findings provide a preliminary comparison of the architectures and warrant further evaluation using patient-level data and external validation.

In future studies, the validation of models on multi-center and independent external datasets should be a priority in terms of assessing the generalizability of the results. In addition, 3D CNN or 2.5D approaches that take into account the volumetric structure of the nodule, different transfer learning architectures, and attention mechanisms could be compared. From a clinical perspective, optimizing the decision threshold particularly with the aim of reducing false negatives, experimenting with cost-sensitive learning approaches, and examining the sensitivity–specificity trade-off across different clinical scenarios would be beneficial. At the final stage, prospective studies in which model outputs are compared with radiologist evaluations could more reliably reveal the contribution of such systems to real clinical decision-support processes.

References

- [1] National Lung Screening Trial Research Team. “Reduced lung-cancer mortality with low-dose computed tomographic screening”, *The New England Journal of Medicine* 365(5) (2011) 395–409; doi: 10.1056/NEJMoa1102873.
- [2] MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M, Rubin GD, Schaefer-Prokop CM, Travis WD, Van Schil PE, Bankier AA. “Guidelines for management of

- incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017", *Radiology* 284(1) (2017) 228–243; doi: 10.1148/radiol.2017161659.
- [3] Armato SG III, McNitt-Gray MF, Reeves AP, Meyer CR, McLennan G, Aberle DR, Kazerooni EA, MacMahon H, van Beek EJR, Yankelevitz D, Hoffman EA, Henschke CI, Roberts RY, Brown MS, Engelmann RM, Pais RC, Piker CW, Qing D, Kocherginsky M, ... Clarke LP. "The Lung Image Database Consortium (LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans", *Academic Radiology* 14(11) (2007) 1409–1421; doi: 10.1016/j.acra.2007.07.008.
- [4] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. "A survey on deep learning in medical image analysis", *Medical Image Analysis* 42 (2017) 60–88; doi: 10.1016/j.media.2017.07.005.
- [5] Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis", *The Lancet Digital Health* 1(6) (2019) e271–e297; doi: 10.1016/S2589-7500(19)30123-2.
- [6] Li D, Vilmun BM, Carlsen JF, Albrecht-Beste E, Lauridsen CA, Nielsen MB, Lindskov Hansen K. "The performance of deep learning algorithms on automatic pulmonary nodule detection and classification tested on different datasets that are not derived from LIDC-IDRI: A systematic review", *Diagnostics* 9(4) (2019) 207; doi: 10.3390/diagnostics9040207.
- [7] Masood A, Yang P, Sheng B, Li H, Li P, Qin J, Lanfranchi V, Kim J, Feng DD. "Cloud-based automated clinical decision support system for detection and diagnosis of lung cancer in chest CT", *IEEE Journal of Translational Engineering in Health and Medicine* 8 (2020) 1–13; doi: 10.1109/JTEHM.2019.2955458.
- [8] Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, van Riel SJ, Wille MMW, Naqibullah M, Sánchez CI, van Ginneken B. "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks", *IEEE Transactions on Medical Imaging* 35(5) (2016) 1160–1169; doi: 10.1109/TMI.2016.2536809.
- [9] Dou Q, Chen H, Yu L, Qin J, Heng PA. "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection", *IEEE Transactions on Biomedical Engineering* 64(7) (2017) 1558–1567; doi: 10.1109/TBME.2016.2613502.
- [10] Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H. "Automated lung nodule detection and classification using deep learning combined with multiple strategies", *Sensors* 19(17) (2019) 3722; doi: 10.3390/s19173722.
- [11] Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, Zang Y, Tian J. "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification", *Pattern Recognition* 61 (2017) 663–673; doi: 10.1016/j.patcog.2016.05.029.
- [12] Xie Y, Xia Y, Zhang J, Song Y, Feng DD, Fulham M, Cai W. "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT", *IEEE Transactions on Medical Imaging* 38(4) (2019) 991–1004; doi: 10.1109/TMI.2018.2876510.
- [13] Mastouri R, Khelifa N, Neji H, Hantous-Zannad S. "A bilinear convolutional neural network for lung nodules classification on CT images", *International Journal of Computer Assisted Radiology and Surgery* 16(1) (2021) 91–101; doi: 10.1007/s11548-020-02283-z.
- [14] Zhu W, Liu C, Fan W, Xie X. "DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification", In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), (2018) 673–681; doi: 10.1109/WACV.2018.00079.
- [15] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Ettemadi M, Ye W, Corrado G, Naidich DP, Shetty S. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography", *Nature Medicine* 25(6) (2019) 954–961; doi: 10.1038/s41591-019-0447-x.
- [16] Venkadesh KV, Setio AAA, Schreuder A, Scholten ET, Chung K, Wille MMW, Saghir Z, van Ginneken B, Prokop M, Jacobs C. "Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT", *Radiology* 300(2) (2021) 438–447; doi: 10.1148/radiol.2021204433.
- [17] Canayaz M, Şehribanoğlu S, Özgökçe M, Akıncı MB. "A comprehensive exploration of deep learning approaches for pulmonary nodule classification and segmentation in chest CT images", *Neural Computing and Applications* 36 (2024) 7245–7264; doi: 10.1007/s00521-024-09457-9.
- [18] Li D, Yuan S, Yao G. "Classification of lung nodules based on the DCA-Xception network", *Journal of X-Ray Science and Technology* 30(5) (2022) 993–1008; doi: 10.3233/XST-221219.
- [19] Kaggle, Lung Nodule Dataset [Online]. Available: <https://www.kaggle.com/datasets/ucimachinelearning/lung-nodule-dataset>.
- [20] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. "Densely connected convolutional networks", In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017) 4700–4708.
- [21] Chollet F. "Xception: Deep learning with depthwise separable convolutions", In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017) 1251–1258.
- [22] Tharwat A. "Classification assessment methods", *Applied Computing and Informatics* 17(1) (2021) 168–192.
- [23] Zhang Q, Wang H, Yoon SW, Won D, Srihari K. "Lung nodule diagnosis on 3D computed tomography images using deep convolutional neural networks", *Procedia Manufacturing* 39 (2019) 363–370.