POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Applied AI in Biomedicine

Laurea Magistrale in Biomedical Engineering

Author: Martina Frageri, Filippo Manini, Laura Moiana

Advisor: Valentina Corino

Co-advisor: Francesca Lo Iacono

Academic year: 2024-2025

1. Introduction

Lung cancer (LC) is the primary cause of cancer-related mortality worldwide and the second most frequently diagnosed cancer globally [1]. Due to the lack of early symptoms, patients often miss the optimal treatment window, making early screening crucial for the prevention and management of lung cancer. The 10-year relative survival rate can reach up to 88% if treatment is administered in time. Pulmonary nodules are clinically relevant as they may represent the initial manifestation of lung cancer. Pulmonary nodules are generally defined as spherical lung opacities or irregular lesions ranging from 3 to 30 mm in size, which can appear as single entities or in multiples [2]. They exhibit diverse characteristics such as quantity (single or multiple), size, shape (regular or irregular), margins (smooth, lobulated, or spiculated), location (well-defined, near the pleura, or adjacent to blood vessels), and density (solid, part-solid, or non-solid).
Visual analysis of medical images is an efficient method to investigate lung tissue, identify nodules, and classify the stages of lung cancer. Various imaging techniques are employed for lung cancer screening [3], including positron emission tomography (PET), computed tomography (CT), ultrasound, chest radiography (X-ray), and magnetic resonance imaging (MRI). However, identifying malignant cells can be challenging due to variations in scan intensity and anatomical structure, often leading to potential misinterpretations by medical professionals [4]. The variability and unpredictability of pulmonary nodules further complicate accurate detection and diagnosis. In recent years, computer-aided diagnosis (CAD) systems have gained popularity as tools to assist radiologists in the accurate diagnosis of lung cancer. Specifically, machine learning (ML) and deep learning (DL) methods have shown great potential in supporting radiologists and clinicians by enhancing nodule detection, classification and segmentation. These AI-driven systems hold the potential to improve cancer diagnosis and prognosis, reduce misclassifications, lower error rates, and provide high-quality imaging analysis.

2. Materials and methods

2.1. Dataset description:

The dataset consists of 2,363 pairs of images in NRRD (Nearly Raw Raster Data) format, where each pair includes a full-slice image and a zoomed-in image of a lung nodule. The full-slice image is a 2D CT scan slice showing the largest visible nodule, while the nodule image is

a focused view of the nodule. Each image pair is annotated with a malignancy score label, ranging from 1 to 5, indicating the severity of the tumor. The distribution of image pairs across the malignancy classes is summarized in Table 1.

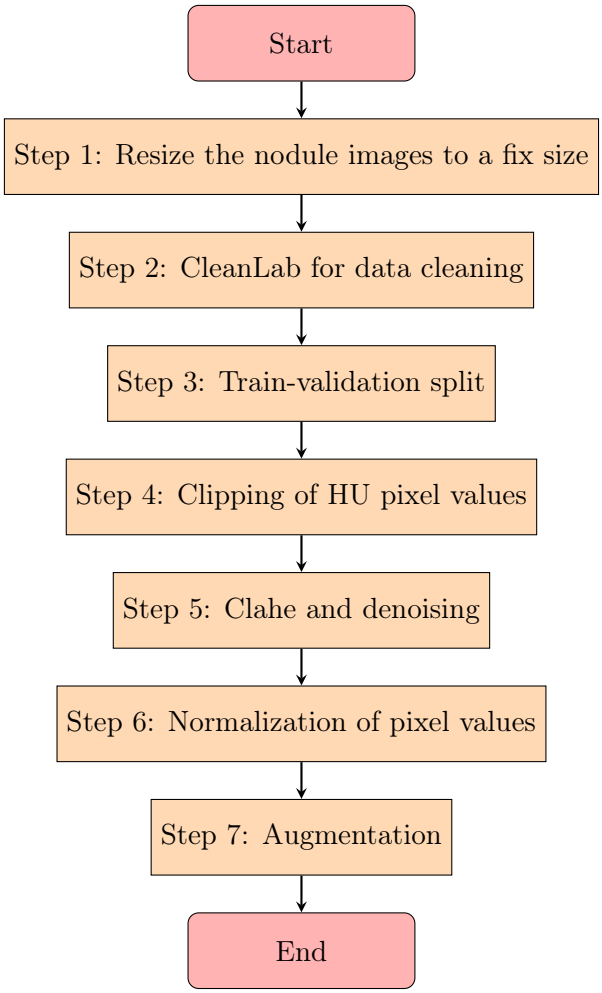
Table 1: Number of images for each class.

Tumor Class	Nr. Images
Class 1	244
Class 2	457
Class 3	1092
Class 4	418
Class 5	152

The full-slice images have a fixed resolution of 512x512 pixels, whereas the nodule images vary in size, with dimensions ranging from 45x44 to 108x124 pixels. The goal of this project is to assign the tumor malignancy to each image performing both 5-class classification and binary classification, where classes 1,2,3 are considered benign and classes 4,5 are considered malignant. By developing separated models for the two types of images, their performance can also be compared, to understand which type of image is more suitable for the tumor malignancy classification.

2.2. Preprocessing:

The steps of the preprocessing phase are summarized in the following flowchart.



To standardize the dataset, all nodule images were resized to a fixed dimension of 64x64 pixels. This resizing step ensures consistency across the dataset and allows the model to process images with uniform dimensions. After resizing, the entire dataset (both full and nodule images) was analyzed using the Cleanlab [5] , an open-source Python library, for automated data cleaning and quality assessment. Cleanlab is specifically designed to identify issues in noisy datasets, such as mislabeled examples, overlapping classes, and other inconsistencies that can adversely impact model performance. Using CleanLab on the full slice dataset, 45 exact duplicate sets and 16 near duplicate sets were identified. Among these, the labels of each set of duplicates were checked. If the images had the same label, only one of them was retained, whereas if the images had different labels, both were removed. So, 87 images were removed obtaining a cleaned dataset of 2276 images.
The dataset was then split into training (80%) and validation (20%) sets, maintaining class balance for proper model evaluation and to mitigate

overfitting. All the images were then analyzed considering the pixel intensity values which, in the dataset, are expressed in Hounsfield Units (HU), a quantitative scale commonly used in CT scans to describe radiodensity. On this scale, denser tissues, which absorb more of the X-ray beam, have positive HU values and appear bright in the images, whereas less dense tissues, which absorb less of the X-ray beam, have negative values and appear darker. The radiodensity of a tissue is directly proportional to its physical density, as denser tissues cause greater attenuation of the X-ray beam. A visual representation of the Hounsfield scale is shown in Figure 1. By definition, water is assigned to a value of 0 HU, and air to -1000 HU. For other materials, bone can reach values as high as +1000 HU, denser bones like those in the cochlea may reach up to +2000 HU, and metals such as steel or silver can exceed +3000 HU [6].

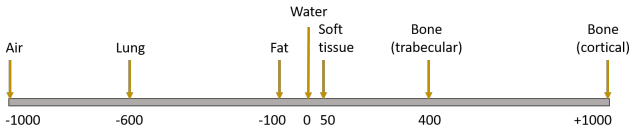


Figure 1: The Hounsfield scale of CT numbers

Considering both full slice and nodule datasets, the HU values were clipped between a minimum value of -1000 and a maximum value of 400, to focus on the range of pixel intensities that are most relevant for lung and soft tissue analysis. After clipping, contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to enhances the local contrast of the images by equalizing the histogram in small regions, which helps preserve finer details without amplifying noise. Additionally, a bilateral filter was applied to reduce noise while preserving edges, maintaining clear boundaries between different tissue types. Then, the pixel values were standardized between 0 and 1 to ensuring that all pixel values are on the same scale, which helps the model converge more efficiently during training.
Finally, as last preprocessing step, data Augmentation [7] was applied to the training set, to improve model robustness and reduce overfitting. Images were augmented with random horizontal and vertical flips, as well as random rotations. More complex augmentations were tested but yielded poor results.

To address class inbalance in the dataset, particularly considering the over-represented class 3, the class weights were computed and normalized. This was done using the *compute_class_weights* function from scikit-learn, which calculates weights inversely proportional to class frequencies. The weights were then normalized to ensure that their sum equals one. By incorporating these normalized weights during training, the penalty for misclassifying minority class instances was increased while it was reduced for the majority class instances. This approach helped to balance the model's attention across all classes, resulting in improved classification performance and better generalization.

2.3. Experiments:

All the experiments were performed on the 5-class classification , as it is the more complex task. Only the best model for each type of image (full-slice and nodule) was later used to perform binary classification.

Full_slice experiments: Various models were designed and adapted for full_slice image classification, starting with a basic CNN architecture as a baseline. The initial model was a simple convolutional neural network (CNN), incorporating residual blocks and attention mechanisms. While this model produced promising results for nodule images, it struggled to perform well on full slice images, with the classification results appearing almost random. This is probably due to the large size of the full_slice images and the small size of the nodules within them. Since the nodule occupies only a small portion of the image, it represents a significant challenge for the network to accurately classify the image based on such a limited region. For these reasons new approaches were applied to improve the performances on the full_slice dataset. One strategy involved the use of Autoencoders to learn better the representation of the images before applying classification. However, this approach did not lead to significant improvements in performance. Indeed, when analyzing the latent space using techniques such as t-SNA, no clear, distinct groups were found for the different classes. Next, transfer learning techniques were explored, using pre-trained models on larger, more general-

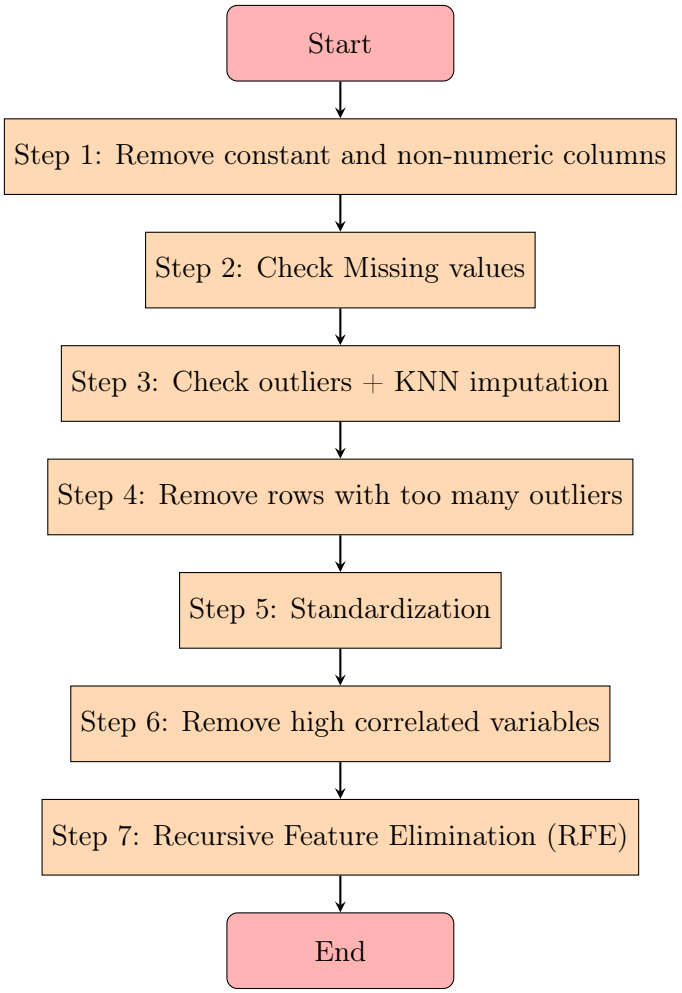
ized datasets like ImageNet, to mitigate the issue of a small dataset. This approach showed some improvements, leading to the conclusion that using pre-trained models can help in the classification task. However, the degree of improvement was limited, leading to the exploration of alternative approaches. Recent researches have demonstrated that pretrained models derived from medical source databases can outperform pretrained models from ImageNet [8] when dealing with medical images, showcasing the potential for achieving better performance with domain-specific data. So, the use of RadImagenet [9] was explored. It consists of a an open radiologic deep learning research dataset that contains a vast collection of 1.35 million medical images, covering CT, MRI and US modalities, across 11 distinct anatomic regions. Available pretrained models on RadImagenet, including InceptionV3, ResNet50, DenseNet121, and InceptionResNetV2, were utilized for transfer learning. These models, having been pre-trained specifically on medical imaging data, were better suited to the task at hand. As a result, significant improvements in performance were observed.

As last experiment, a novel approach was introduced to address the challenge of detecting the nodule, which is very small compared to the rest of the image. This new method involved providing the network with coordinates of the nodule center, computed through template matching using the cv2.matchTemplate() function. This technique helped to localize the nodule's position in the image more effectively, improving the model's ability to focus on the relevant regions. This adjustment led to the transformation of the problem from a simple classification task into a multi-task learning problem, where the model shares a common backbone — InceptionResNetV2 - but has two specialized heads designed to handle different tasks. The first head is dedicated to classification, using a softmax activation function and a Categorical CrossEntropy loss function with class weights to address the class imbalance in the data. This allows the network to learn to classify images based on the nodule's presence and category. The second head is designed for regression, predicting the normalized coordinates of the nodule's center within the image. These coordinates are scaled between 0 and 1, which helps mitigate the risk of divergent behavior during training due to the large range of potential coordinate values. To evaluate this task, the Mean Squared Error (MSE) loss function was used, which is ideal for regression tasks involving continuous values. One challenge we encountered during training was the potential imbalance between the two tasks. Since the scale of the classification loss and the regression loss can differ significantly, the model could have become biased toward one task, negatively affecting performance. To mitigate this, we implemented a dynamic loss weighting callback, which adjusts the contribution of each loss term (classification and regression) based on their relative magnitudes at the end of each epoch. By updating the weights dynamically, the callback ensures that neither loss term dominates the training process, helping to balance the focus of the model on both tasks and improving overall performance.

Nodule experiments: As with the full_slice images, the basic CNN architecture was initially used for the nodule images. However, this time it produced promising results, probably because the nodule occupies a much larger portion of the image, and so the network could more easily focus on the region of interest, leading to better feature extraction and classification accuracy. Given that the zoomed-in images contained fewer irrelevant details and focused more directly on the nodule, a different approach was subsequently explored. Specifically, radiomic features were extracted using pyradiomics [10], an open-source python package for the extraction of Radiomics features from medical imaging. The features extracted by pyradiomics include first-order statistics (such as mean and standard deviation), geometric properties (shape-based), and texture measures (such as GLCM, GLRLM, and GLSZM), which analyze the spatial relationships and intensity variations between pixels. These features provide detailed descriptions of the distribution, shape, and texture of regions of interest in medical images, aiding in automatic image analysis, such as for nodules or tumors. However, the exact tumor segmentation was not provided so, reduce the background in these images become a crucial step. To mitigate this, the images were cropped

to 70% of the initial image, focusing on the central region where the nodule is located, in an attempt to reduce the amount of background while retaining the nodule itself. However, since this method of cropping was done without precise segmentation, it only provided an approximation. Ideally, segmentation would allow the exact region of the tumor to be isolated, leading to more accurate feature extraction. This method leads to the extraction of 487 features, and a preprocessing to select the most relevant ones was applied.

First, constant columns were removed, reducing the feature set by 58 variables. Additionally, non-numeric columns, which included image identifiers, center coordinates and image dimensions, were identified and excluded as they did not provide meaningful information for classification, resulting in the removal of 7 more features. Then, missing values (Nan) were checked but none were found. Outliers were then identified using the z-score method, where values with an absolute z-score greater than 3 were considered outliers. These were replaced with NaN, resulting in 3843 missing values in the training set and 916 in the validation set. Rows in the training set containing more than 1/3 of their feature values missing were removed, which led to the deletion of 17 rows. The remaining missing values were imputed using KNN imputation, with k=5). The dataset was then normalized using MinMaxScaler to ensure consistency across the dataset. Following, highly correlated variables, with a correlation coefficient greater than 0.8, were eliminated, keeping the variable with the lower correlation to the all the remaining features. This step reduced the feature set by 100 variables. Finally, a Recursive Feature Elimination (RFE) process was performed using a Random Forest Classifier. This method involved the removal of one feature at time and the evaluation of the resulting model using k-fold-CV with k=3. The goal was to identify the optimal subset of features that maximize the balanced accuracy. This process resulted in the selection of 18 features, which were found to be the most relevant for classification. The described pre-processing steps are summarized in the following flowchart:



Once the best features were extracted, they were initially used to train various machine learning models, including Support Vector Machine (SVM), Random Forest and Neural Network (NN). After evaluating their performance, the NN was chosen as the preferred model due to its superior accuracy. This approach with radiomic features yielded promising results in the classification task, and for this reason, it was further developed. Indeed, to improve the performance, an integrated approach was developed that combined both the radiomic features and the nodule images in a single NN. This model was designed with two distinct input paths: one for the 18 radiomic features and one for the nodule images. The architecture consisted of two separate branches, each initially processing its respective input independently. The "image branch" was composed of five convolutional layers. These layers alternated between basic convolutional layers (followed by batch normalization, max pooling, and squeeze-and-excitation (SE) blocks) and residual blocks (followed again by max pooling and SE blocks). To capture fea-

tures at multiple scales, the second convolutional layer was replaced with an inception block, allowing the network to extract features at different levels of abstraction. After these five layers, the global average pooling (GAP) was applied to flatten the inputs, followed by a fully connected dense layer with dropout to prevent overfitting. On the other hand, the "radiomic features branch" consisted of three dense layers with ReLU activation functions, batch normalization, and dropout to enhance generalization. A skip connection was added between the second and third layers to ensure a more effective flow of information through the network. Before combining the two branches, batch normalization was applied to each to standardize the activations and improve the stability of the network. Then the two branches were concatenated and passed through two final dense layers, followed by a dropout, to combine the information from both sources. At the end, the output layer produced the final classification predictions. This integrated model demonstrated significantly improved performance, highlighting the advantage of merging complementary information from both the quantitative radiomic features and the visual data from the nodule images. Using both types of data, the model was able to achieve better generalization and classification accuracy compared to using either the features or the images alone.

2.4. Training:

During training, several strategies were implemented to optimize model performance and improve generalization. AdamW optimizer was chosen to offer a good balance between convergence speed and stability. The learning rate was initially set to 1e-3 or 1e-4 depending on the network and dataset. A learning rate schedule using ReduceOnPlateau was introduced to dynamically adjust the learning rate when the model reaches a plateaus during training. This helped to avoid overfitting and improve generalization. Another callback used is the early stopping, which stops the training process when the validation loss doesn't improve for a certain number of epochs (Patience) during the training. This technique helps to save computation time and to preventing overfitting. The loss function was chosen based on the classification task: for multi-class classification, categorical cross-entropy was used, while binary cross-entropy was applied for binary tasks. To train the model, the mini-batch gradient descent method was chosen, which involved dividing the dataset into batches. For the nodule images, a batch size of 64 was selected. However, for the full-slice images, the batch size was reduced to 16 due to the significantly larger size of these images, which increased computational requirements. Finally, to assess the model's ability to generalize, performances were evaluated on the validation set and the confusion matrix was performed as visual assessment to ensure the model was not over-predicting the most common class but was generalizing well across all the five classes.

3. Results

Using the full-slice images, several experiments were conducted, including basic CNNs and autoencoders but none of them produced good results, with accuracy over the 5 classes around 28%, indicating the model was performing randomly. Then, pre-trained models on ImageNet were employed to address the limitation of the small dataset. They showed little improvements, with a categorical accuracy of arounf 43%, leading to the conclusion that pre-trained models are useful for the classification task. Ultimately, pre-trained models on RadImagenet were employed, which showed further improvements (even if still not ideal) with categotical accuracy reaching approximately 50%, demonstrating that the model was no longer behaving randomly, as we can also see looking at the high performance on the binary task (83% of binary accuracy). The last experiment - the center regression - didn't show improvements from the TL on RadImagenet. For detailed metrics, please refer to the Table 2. Looking at the confidence of each prediction in Figure 2 (based on the output of the last layer, either softmax for categorical tasks or sigmoid for binary ones), it is clear that correct predictions, as expected, have higher confidence compared to incorrect ones across all the classes. However, when comparing correct confidence levels between different classes, it becomes apparent that all classes show high confidence (>70%), but classes 1 and 2 have lower confidence with greater variance.

Table 2: Full_slice Models' metrics on validation set for the Categorical classification (on the left) and the Binary classification (on the right). Best results in **bold**.

Model	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Basic CNN	42.11	27.65	42.11	-	-	-
Autoencoders	-	-	-	-	-	-
Center regression	45.61	46.2	45.61	-	-	-
TL Imagenet	43.42	39.88	43.42	-	-	-
TL RadImagenet	**50.00**	**47.84**	**50.00**	**83.99**	**84.32**	**83.99**

This observation is further reflected in the confusion matrix, where these two classes emerge as the most challenging to predict accurately.

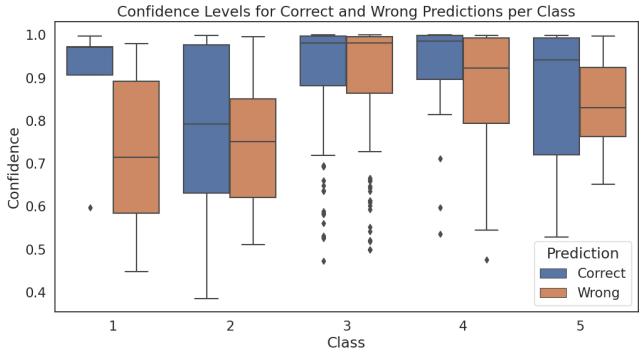


Figure 2: Confidence interval over the 5 classes for the full_slice images

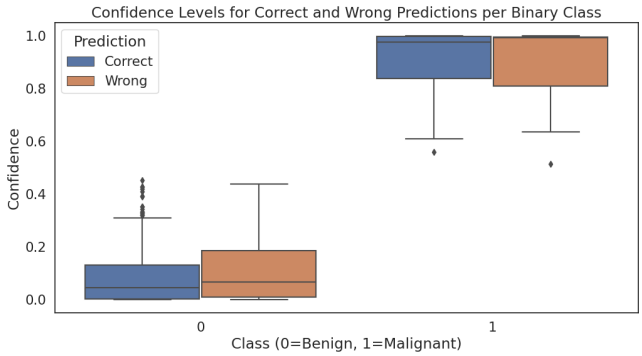


Figure 3: Confidence interval over the binary classes for the full_slice images

For the nodule images, the basic CNN already yielded promising results, achieving approximately 55% of accuracy over the 5 classes. Additionally, using only features extracted with PyRadiomics also produced promising outcomes, with a categorical accuracy of around 50%. By combining both approaches in a unique NN, the best results were obtained, with an accuracy of approximately 61%. For detailed metrics, please refer to Table 3. Additionally, analyzing the confidence levels in Figure 4 it is clear that, as with the full-slice model, correct predictions exhibit higher confidence. However, when comparing confidence between classes, is evident that class 3 has a much higher confidence (a median around 90%), compared to the other classes which show median confidence levels around 50%, 40%, 50%, and 70% for classes 1, 2, 4, and 5, respectively. From this, we can also infer that class 2 is the most challenging to predict, followed by classes 1, 4, and 5, while class 3 is relatively easier. This observation is also reflected in the confusion matrix.

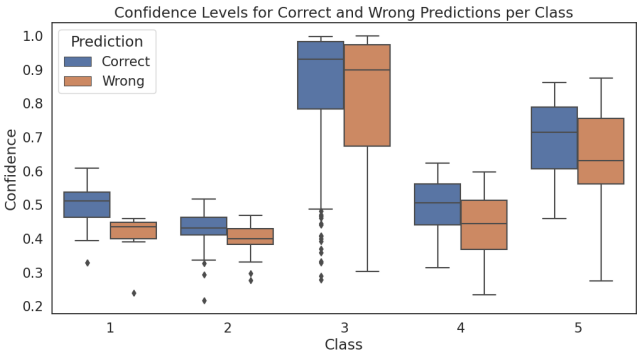


Figure 4: Confidence interval over the 5 classes for the nodule images

Table 3: Nodule Models' metrics on validation set for the Categorical classification (on the left) and the Binary classification (on the right). Best results in **bold**.

Model	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Basic CNN	55.48	53.66	55.48	-	-	-
Features	49.56	52.51	49.56	-	-	-
Features + images	**61.84**	**61.97**	**61.84**	**89.47**	**89.25**	**89.47**

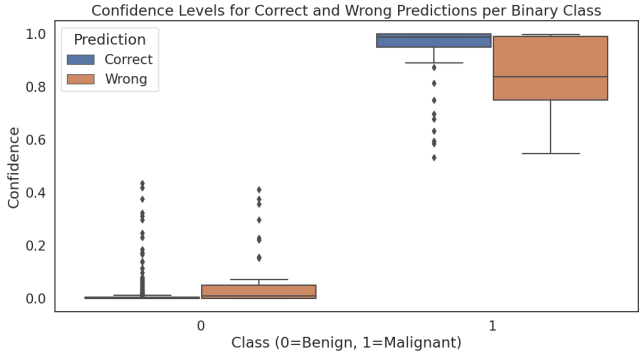


Figure 5: Confidence interval over the binary classes for the nodule images

4. Discussion

The results demonstrate a significant performance difference between models trained on full-slice images and those trained on nodule images. Indeed, models of nodule images achieved higher accuracy, while the ones of full_slice images struggled, even after extensive experimentation. This is probably due to the complexity of the full_slice images, with nodules occupy only a small portion of the image, with a lot of surrounding anatomical details that may confuse the model. Although autoencoders and transfer learning from ImageNet were attempted, these methods struggled probably because the features learned from natural images do not translate well to medical images. This highlights the need for domain-specific knowledge in medical image analysis. In contrast, nodule images focus specifically on the region of interest, eliminating much of the irrelevant background. This simplification allowed the models to better capture key features associated with tumor classification. The use of PyRadiomics further improved results by providing quantitative features that capture details, particularly benefiting the prediction of minority classes. Moreover, the introduction of the radiomic features, reduced the amount of overfitting, leading to better generalizable models. When looking at the confidence levels of the predictions, some interesting differences emerge between the two models. For the full-slice model, the correct prediction confidence values are generally higher across all classes, indicating that the model is more confident in its predictions, regardless of the final accuracy. In contrast, the model based on nodule images shows a significantly higher confidence for class 3 compared to the others, which is also the most represented class in the dataset. This suggests that the model is particularly confident when predicting this class, while the other classes have much lower confidence levels (around 40-70%). This could indicate that the nodule model is less balanced in its predictions, being influenced by the class distribution. So, even if the nodule image model achieves higher accuracy, the full_slice model demonstrates more consistent confidence across classes. This highlights that both models have specific strengths and weaknesses.

Indeed, for both images type, several challenges remain. First, the dataset was relatively small, which limits the amount of training data available for the models to learn robust representations. Moreover, the dataset was also imbalanced, with certain tumor classes being underrepresented, which led to biased models which predict more the majority class. Despite attempts to counter this imbalance, such as using weighted losses or data augmentation, overfitting was still a prominent issue. So, the models memorize patterns in the training data without generalizing well to new samples. Another limitation was the lack of segmentation data. Indeed, without explicit segmentations for the nodules, the models had to rely only on the raw images to find the region of interest, which is very difficult expecially in full-slice images due to the presence of irrelevant anatomical de-

tails. In the case of nodule images, the lack of segmentation limited the extraction of accurate radiomic features. Therefore, implementing segmentation could improve model performance for both full-slice and nodule images. In conclusion, the difference in performance between the 2 models highlights the difficulties in using raw medical images for classification tasks. Probably, addressing these limitations, the use of larger, balanced datasets and the incorporation of segmentations, could lead to an improvement in the model's performance in future experiments.

5. Conclusion

This study explored the performance of deep learning models in classifying two types of lung CT images: full_slice images and nodule-focused images. The results showed that models trained on nodule images outperformed those trained on full_slice images, likely due the fact that nodule images focus directly on the region of interest. Despite the promising results, several limitations remain. The small and imbalanced dataset limited the robustness of the models, and the lack of segmentation made the models struggling to localize nodules. Future works could focus on addressing these limitations by using larger, more balanced datasets and incorporating segmentation data. Overall, this study highlights the complexities of working with medical images in deep learning, underlying the importance of domain-specific techniques and the need for well-focused images that contains the task-relevant features without unnecessary details that may confuse the model.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Quanyang Wu, Yao Huang, Sicong Wang, Linlin Qi, Zewei Zhang, Donghui Hou, Hongjia Li, and Shijun Zhao. Artificial intelligence in lung cancer screening: Detection, classification, prediction, and prognosis. *Cancer Medicine*, TBD:TBD, 2024. Received: 24 November 2023; Revised: 15 March 2024; Accepted: 16 March 2024.

[3] Rabia Javed, Tahir Abbas, Ali Haider Khan, Ali Daud, Amal Bukhari, and Riad Alharbey. Deep learning for lungs cancer detection: a review. *Artificial Intelligence Review*, 57:197, 2024. Accepted: 16 May 2024; Published online: 8 July 2024.

[4] Mohammad A. Thanoon, Mohd Asyraf Zulkifley, Muhammad Ammirrul Atiqi Mohd Zainuri, and Siti Raihanah Abdani. A review of deep learning techniques for lung cancer screening and diagnosis based on ct images. *TBD*, 2024. Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia; System and Control Engineering Department, Ninevah University, Mosul 41002, Iraq; School of Computing Sciences, Universiti Teknologi MARA, Shah Alam 40450, Malaysia.

[5] Cleanlab, https://github.com/cleanlab/cleanlab.

[6] T. D. DenOtter and J. Schubert. *Hounsfield Unit*. StatPearls Publishing, Treasure Island (FL), Mar 2023. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK547721/`.

[7] Keras preprocessing layers: Image augmentation.

[8] ImageNet. Imagenet: A large-scale hierarchical image database. `https://www.image-net.org`. Accessed: 2025-02-10.

[9] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Journal of Digital Imaging*, 35(5):1376–1386, 2022.

[10] PyRadiomics. Pyradiomics documentation. `https://pyradiomics.readthedocs.io/en/latest/`. Accessed: 2025-02-10.