Lung Nodule Classification in CT Images Using Deep Learning

Anna Simeone, Pasquale Serrao, Carlotta Pecchiari, Marta Zecchini

245744, 250803, 233381, 246790

February 12, 2025

Contents

1 Introduction

Lung cancer is one of the deadliest malignancies, responsible for 18% of cancer-related deaths in 2020 [1]. Despite treatment advances, prognosis remains poor due to **late-stage diagnosis**. Many cases are asymptomatic early on, delaying detection until the disease becomes less treatable. Early and accurate diagnosis is crucial for improving survival rates. Computed tomography (CT) scans are the gold standard for screening, detecting small nodules indicative of malignancy. However, manual interpretation is time-consuming and prone to variability. To address these challenges, integrating **artificial intelligence (AI)** and **deep learning** into lung cancer diagnosis offers a promising solution.

Deep learning has demonstrated remarkable performance in medical image analysis, outperforming traditional machine learning methods in feature extraction and classification tasks. AI-driven models can autonomously analyze CT scans, detect lung nodules, and assess malignancy risk, thereby enhancing the accuracy, efficiency, and consistency of cancer screening programs.

This project develops a deep learning-based lung nodule classification system for both binary (benign vs. malignant) and multi-class (malignancy score 1–5) classification. The model will include a confidence estimation to assess prediction reliability. A key focus is comparing full-slice CT scans, which provide a broad anatomical context, with zoomed nodule images, which highlight the lesion. This analysis aims to identify the most effective approach for AI-assisted lung cancer screening, with the goal of improving early detection and supporting radiologists in clinical decision-making.

2 Materials and Methods

2.1 Dataset

The dataset comprises computed tomography (CT) scan images of lung nodules from approximately 2,400 patients.
For each patient, the dataset includes:

- **Full-Slice CT Image:** The entire CT scan slice containing the lung nodule. This image provides a broader anatomical context, capturing surrounding lung tissue and structures that may be relevant for classification.
- **Zoomed-Slice Image:** A cropped region centered on the detected lung nodule, isolating the lesion from the surrounding anatomy. This focused view reduces background noise and enhances the visibility of critical features.

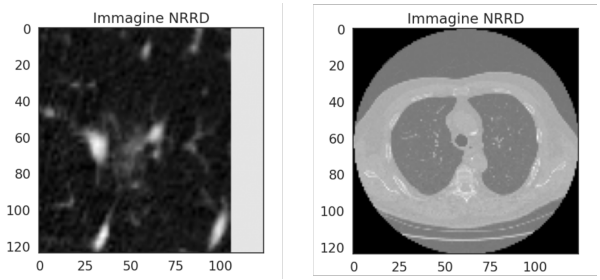

Figure 1: Example of a nodule (left) and full-slice image (right) of the same patient.

For each patient, a malignancy score from 1 to 5 is assigned. For multi-class classification task, the model should predict one of the five malignancy levels. For binary classification task, labels are grouped as follows:

- **Benign:** scores 1, 2 and 3
- **Malignant:** scores 4 and 5

In this case the model should be able to predict if the cancer is malignant or not.

2.2 Challenges

There are several challenges in our work that need to be addressed, with the most significant being:

- **Data Scarcity:** The dataset available for model implementation was limited, which can lead to *poor generalization and biased predictions*. Neural networks require large amounts of data to learn meaningful representations, and insufficient data may hinder their performance.
- **Data Imbalance:** In the binary classification task, the benign class was *overrepresented*, with 1,793 benign samples compared to only 570 malignant ones. Similarly, in the multi-class classification task, class 3 had significantly more samples than all other classes (see figure 2).

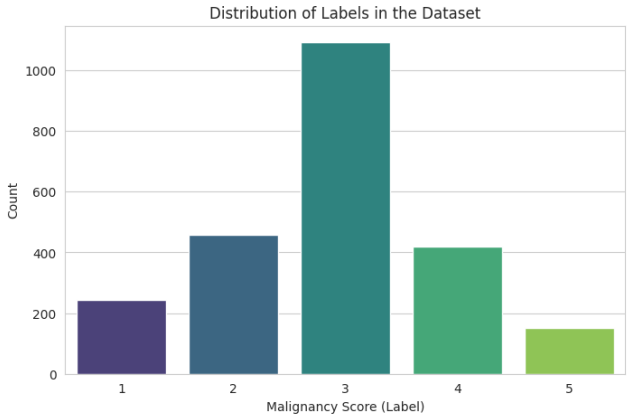

Figure 2: Label distribution in the dataset.

This imbalance can bias the model toward the majority class, leading to a high false-negative rate and poor minority class detection. Addressing this issue is crucial to improving model performance.

- **Generalization Capability:** When dealing with data scarcity and class imbalance, a major challenge is avoiding *overfitting*. With limited and imbalanced data, the model may memorize patterns specific to the training set rather than learning generalizable features. Overfitting leads to high accuracy on training data but poor performance on unseen validation and test samples. To address it is essential to improve generalization and prevent the model from becoming overly dependent on specific training patterns.

2.3 Assumptions

The initial assumptions are:

- **Problem is well-defined**: the task involves a finite set of classes (labels), and each input image belongs to one of these classes.
- **Accurate class labels:** the labels are correctly annotated by medical experts, so there are no significant label noise or errors that could mislead the model.

2.4 Data inspection and pre-processing

After inspecting the dataset and confirming the absence of unlabeled images, we proceeded with a semantic analysis. To achieve this, we first trained an **autoencoder** to extract the main features of the images, projecting them into the latent space. Once trained, the decoder was discarded, and we used the encoder to visualize the images in this latent space. Subsequently, we applied **Principal Component Analysis (PCA)** and **Isolation Forest** to the transformed image representations to identify potential semantic outliers. Upon examining the images classified as outliers, both in the full-slice and nodule cases, we found no significant differences compared to the rest of the dataset. As a result, we concluded that these instances were likely just particular cases rather than actual anomalies, and thus decided not to remove them.

For image preprocessing, we normalized the data using **Min-Max normalization** to standardize the input and improve both model performance and convergence speed. Subsequently, we applied **Sigmoidal Contrast** (see figure 3), which enhances contrast using a sigmoid function, as shown below:

$$S(x) = \frac{1}{1 + e^{-\text{gain}\cdot(x-\text{cutoff})}}$$

where $S(x)$ is the enhanced pixel value at position x, **cutoff** represents the gray value around which contrast adjustment occurs (typically the mean pixel value), **gain** controls the steepness of the sigmoid curve, regulating how quickly the contrast is enhanced.

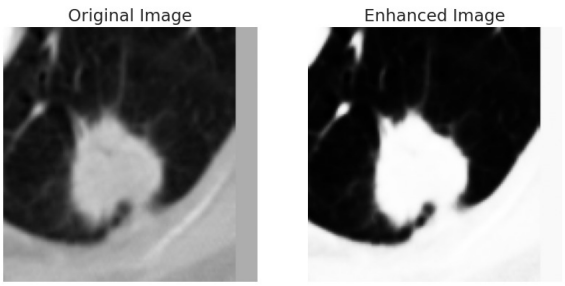

Figure 3: Example of a nodule image before and after applying Sigmoidal Contrast.

This type of preprocessing is particularly useful for medical images such as CT scans, where subtle details need to be enhanced without distorting essential structures.

Additionally, for the nodule images we applied a center crop (96×96) to remove recurrent gray bands present in the images. Finally, to maintain consistency with the rest of the dataset, we resized

the cropped images back to 128×128 before feeding them to the model.

The dataset was divided into *training* (70%), *validation* (15%), and *test* set (15%) using stratification.

To address **class imbalance**, we explored different strategies. Initially, we applied a **weighted loss function**, assigning weights inversely proportional to class frequency to give more importance to errors on underrepresented classes.

However, this approach was later discarded in favor of directly balancing the dataset, which yielded better performances. At first, we **upsampled** the minority classes using augmentation to match the most frequent class. However, this introduced a potential bias, as the model could learn to distinguish the majority class simply due to its lack of augmentation. To mitigate this issue, we adjusted the upsampling strategy by increasing the number of samples in each class until they all reached the size of the most represented class in the training set, plus an additional 200 samples. This ensured that even the majority class underwent the same augmentation process, leading to a more balanced and unbiased dataset for training.

2.4.1 Other attempts

We also experimented with a different image enhancement technique, named **CLAHE** (figure 4), and incorporated some morphological operations as described in the paper [2]. However, we ultimately decided to discard this approach, as it did not lead to improved performances and appeared to confuse the model.

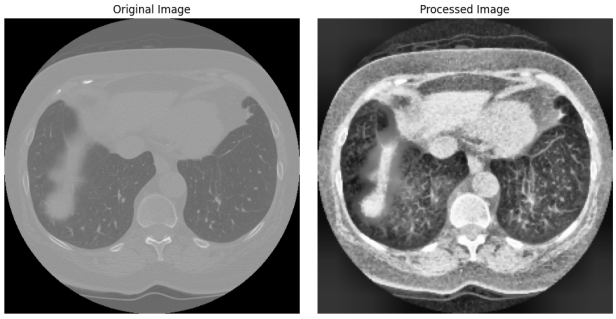

Figure 4: Example of CT scan before and after applying CLAHE and morphological operations.

3 Model Architecture

All the models were trained separately on different image datasets:

1. **Binary classification**: distinguishing between benign and malignant cases.

2. **Multi-class classification**: predicting malignancy scores from 1 to 5.

We explored multiple architectures to evaluate their effectiveness for the specific tasks. We followed two different approaches, one based on the development of a custom **Convolutional Neural Network (CNN)** and the other involving **Transfer Learning (TL)** and **Fine Tuning (FT)**.

The custom CNN was made of four convolutional blocks, where each of them included a convolutional layer, a **batch normalization layer** to stabilize the training, and a max-pooling layer to reduce spatial dimensions.

The second approach utilized pre-trained models, including *MobileNet*, *ResNet50*, and *EfficientNet*, all originally trained on the *ImageNet* dataset. We ultimately chose **EfficientNetB2**, as it consistently delivered the best performance. This model was employed as feature extractor and after the convolutional base of the pre-trained network, a **Global Average Pooling (GAP)** layer was added to obtain a compact and efficient representation of the learned features. To further prevent overfitting, a **Dropout** layer was introduced. The final step involved a Dense layer with *softmax* activation in multi-class classification models, while we used *sigmoid* activation in binary classification models.

4 Model training

For the pre-trained models we adopted a gradual fine-tuning approach to preserve the integrity of the pre-trained weights. Initially, we trained only the top-added layers until convergence. Once the top layers were trained, we selectively fine-tuned a sub-set of layers. The layers closer to the top of the network capture deep semantic features relevant for each specific task, so their fine-tuning helped improving the classification performance. For the training process, we considered the following aspects:

1. **Loss function:** We used *Cross-Entropy* as the loss function, as it is well-suited for classification tasks. Specifically, we applied *Binary Cross-Entropy* for binary classification and *Categorical Cross-Entropy* for the multi-class problem.

2. **Optimizer and learning rate:** We adopted the *Adam* optimizer, known for its efficiency and adaptability in training deep neural networks. A key factor in optimization is the *learning rate*, which controls how much the model updates its weights during training. If the learning rate is too high, the model may struggle to converge, overshooting the optimal solution. Conversely, a very low learning rate can lead to slow training and the risk of getting stuck in a suboptimal local minimum. To further improve generalization and prevent overfitting, we incorporated **weight decay**, a regularization techniques that works by adding a penalty term to the loss function, proportional to the sum of the squared weights, discouraging large weight values.

3. **Batch size:** The choice of batch size plays a crucial role in training dynamics. Larger batch sizes generally lead to more stable and smoother updates, as they provide a more accurate estimate of the gradient. However, they require more memory and may reduce generalization. Due to computational constraints, we selected a batch size of 128, striking a balance between efficient training, stability, and available resources.

4. **Early stopping:** It is a useful callback to train the model. In particular, it continuously monitors the model's performance on a validation set and stops training when improvement ceases and performances begins to degrade. By stopping training once the validation loss stops decreasing, it helps prevent the model from overfitting the training data.

5. **Dynamic data augmentation:** To further improve generalization and mitigate overfitting, we applied *dynamic data augmentation* during training. This technique introduces random transformations to each batch, effectively increasing the diversity of the dataset without requiring additional data collection. The augmentations included random rotations, horizontal and vertical flips and translations. By dynamically modifying input images at each training iteration, the model learns more robust and invariant features, reducing its reliance on specific patterns present in the training set.

4.0.1 Other attempts

In addition to standard augmentation techniques, we experimented with more advanced methods such as *RandAugment* (see figure 5) and *AugMix*, which have shown promise in enhancing model robustness. However, these approaches did not lead to performance improvements. A possible reason is that the transformations applied were too aggressive, potentially distorting key semantic features in the dataset. This may have made it harder for the model to extract meaningful patterns, ultimately hindering rather than enhancing generalization.

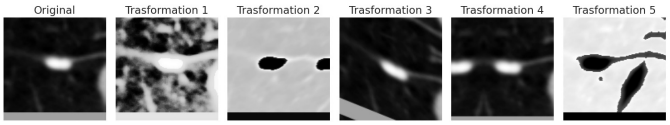

Figure 5: Example of RandAugment applied to a nodule image.

Test Time Augmentation (TTA) was also tested to enhance model robustness. We applied five transformations to each test image and used the most frequent prediction (mode) as the final output for binary classification. This approach was intended to improve prediction reliability; however, it led to a decrease in performance. In most cases, the model ended up predicting only class 0, which was not the desired outcome.

5 Results and discussion

For evaluating model performance and select the best one we used the **F1-score** as the primary metric. Given the class imbalance in our dataset, accuracy alone would not be a reliable indicator, as a model biased toward the majority class could still achieve high accuracy without effectively classifying the minority class. The F1-score, being the harmonic mean of precision and recall, provides a more

balanced assessment, ensuring that both false positives and false negatives are considered, making it more robust in imbalanced scenarios.

The main results are highlighted in the two following tables, which include also the baseline model (CNN):

Table 1: Performance comparison (F1-score on test set) of different models across classification tasks.

Model	**Binary Full-Slice**	**Binary Nodules**	**M-C Full-Slice**	**M-C Nodules**
CNN	56.41%	65.57%	29.43%	34.3%
EfficientNet TL	**69.61%**	77.95%	35.4%	39.36%
EfficientNet FT	66.10%	**81.21%**	**39.49%**	**44.63%**

Table 2: Table of the main results achieved.

Model	Precision	Test Acc	F1 score
Binary_FullSlice	68.30%	72.57%	66.10%
Binary_Nodule	81.14%	82.28%	81.21%
MultiClass_FullSlice	39.06%	42.19%	39.49%
MultiClass_Nodule	46.47%	44.73%	44.63%

5.1 Binary Classification Performance

The results of the binary classification tasks are summarized in Table 1 and Table 2. Among the tested models, the **EfficientNetB2 Fine-Tuned (FT)** achieved the best performance in the Binary Nodules classification task, where it reached an F1-score of 81.21%, significantly outperforming the baseline CNN. Similarly, in the Binary Full-Slice classification, **EfficientNetB2 with Transfer Learning (TL)** approach showed notable improvement with respect to the baseline, obtaining an F1-score of 69.61%.

In terms of overall scores, the Binary Nodules classification exhibited the highest performance, suggesting that the model effectively captured relevant patterns when analyzing localized features, such as nodules. Conversely, the binary model trained on full-slice images performed slightly worse, likely due to the broader context present in the images. The presence of additional *non-relevant anatomical structures* could, indeed, interfere with the model's ability to distinguish between benign and malignant cases.

5.2 Multi-Class Classification Performance

The multi-class classification task proved to be more challenging than the binary tasks, as reflected in the overall lower F1-scores across both models trained on full-slice and nodule images (see Table 1 and Table 2). The best-performing model was **EfficientNetB2 Fine-Tuned (FT)** for both of them, achieving an F1-score of 39.49% for the model trained on full-slice images and 44.63% for the model trained on nodule images.

One possible explanation for the lower performance is the *increased complexity of the task*. Unlike binary classification, where the model only distinguishes between two categories, multi-class classification requires the model to learn **finer-grained distinctions** between multiple classes. This is particularly challenging because the boundaries between malignancy classes aren't well defined, leading to a higher risk of misclassification.

Once again, the model trained on full-slice images appeared less capable than the corresponding one trained on nodule images in performing classification. The presence of extensive background and anatomical variations may have introduced irrelevant contextual information, leading the model to misinterpret key regions. The model trained on nodule images, instead, could focus more specifically on the nodule itself, resulting in relatively better performance.

5.3 Comparison Between Full-Slice and Zoomed-Slice Input

Comparing **full-slice** and **zoomed-slice** images for binary classification, the latter yielded better results. This outcome is expected, as zooming in on the nodule isolates the most relevant features for malignancy prediction, whereas full-slice images include surrounding anatomical structures that may introduce misleading information.

5.3.1 Explainability and Model Confidence Analysis

To further analyze the model's decision-making process, we leveraged explainability techniques, including **Grad-CAM** and **LIME**. These methods helped us visualize which regions of an image contributed most to the predictions and assess the model's reliance on relevant features.

Grad-CAM (Gradient-weighted Class Activation Mapping) allowed us to visualize which regions of an image contributed most to the model's predictions by leveraging gradients from the final convolutional layer, making it particularly useful for our CNN-based model. From the **Grad-CAM** visualizations, we observed that in the full-slice images, the model often distributed its attention across a broad region, sometimes focusing on irrelevant anatomical structures instead of the actual nodule (6). This lack of precise localization likely contributed to misclassifications, particularly in the multi-class classification task, where distinguishing between different malignancy levels requires fine-grained feature extraction.

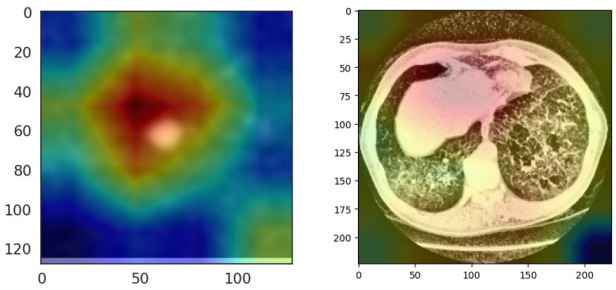

Figure 6: Example of Grad-CAM applied to a nodule (left) and to a full slice image (right) in the multi-class problem.

Conversely, for zoomed-slice images, the attention maps were more concentrated around the nodule, supporting the hypothesis that focusing on the lesion itself improves classification accuracy (6)

To further interpret the model's predictions, we employed **LIME (Local Interpretable Model-Agnostic Explanations)**. LIME approximates a complex deep learning model with a simpler, interpretable surrogate model by perturbing the input data and analyzing the impact of these perturbations on the model's predictions.

From our analysis, LIME provided valuable insights into feature importance at the pixel level especially in the nodule images (7). In the classification, it highlighted key regions within the lesion that strongly influenced the decision, confirming that the model was learning relevant patterns. However, in the full-slice setting, LIME revealed that the model sometimes assigned importance to surrounding anatomical structures rather than the nodule itself. This aligns with our observations from **Grad-CAM**, further supporting the idea that full-slice models may rely on contextual cues rather than lesion-specific features.

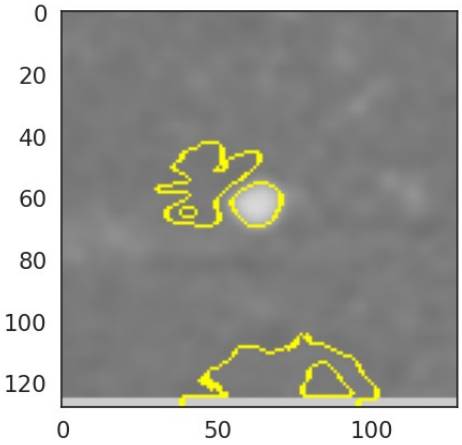

Figure 7: Example of LIME applied to a nodule image in the multi-class problem.

In addition to feature attribution techniques like **Grad-CAM** and **LIME**, we leveraged **Image Retrieval** to gain deeper insights into the model's decision-making process. Image Retrieval works by comparing learned embeddings, allowing us to find images that the model considers similar based on its internal feature representations.

By retrieving the most similar images for a given test sample, we could analyze how the model grouped and classified different cases. This approach provided an intuitive way to verify whether the model was making decisions based on relevant patterns or if it was overly influenced by irrelevant features. In the nodule-based classification, the retrieved images often contained lesions with similar texture, shape, and intensity, suggesting that the model successfully captured meaningful visual features, even though they might belong to different classes. This highlighted how tricky the classification problem might be.

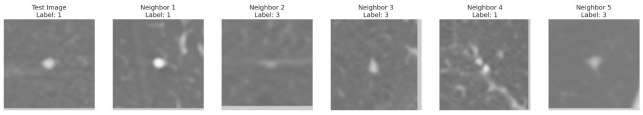

Figure 8: Example of Image Retrieval applied to a nodule image in the multi-class problem.

The **confidence levels** assigned by the model provide valuable insights into how well it distinguishes between different classes and how reliable its predictions are. The following visualizations (9 and 10) illustrate key aspects of model confidence: the **distribution of confidence scores** across predictions and a **comparison between predicted and true confidence per class**.

For brevity, we present the results for the multi-class model on nodules, as the trends observed are consistent across all models. The findings and conclusions drawn from this analysis apply similarly to the binary classification models and the multi-class full-slice model, as they exhibit comparable behavior in terms of confidence distribution.

The first histogram (9) represents the distribution of confidence scores for predicted labels across the test set. The mean confidence value is approximately 0.64, suggesting that, on average, the model exhibits moderate certainty in its predictions. However, the distribution is highly variable, with confidence scores ranging from 0.3 to 1.0.

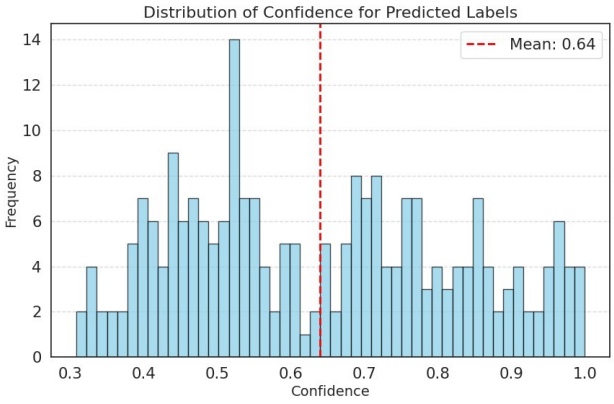

Figure 9: Confidence scores for the predicted labels across the test set.

A significant portion of predictions have confidence values between **0.5 and 0.7**, indicating that the model often makes predictions with medium confidence. There are noticeable peaks at higher confidence levels, suggesting that for some samples the model is highly certain. Conversely, a non-negligible number of predictions exhibit lower confidence scores (below 0.5), meaning the model struggles with certain cases, possibly due to class overlap or insufficient discriminative features.

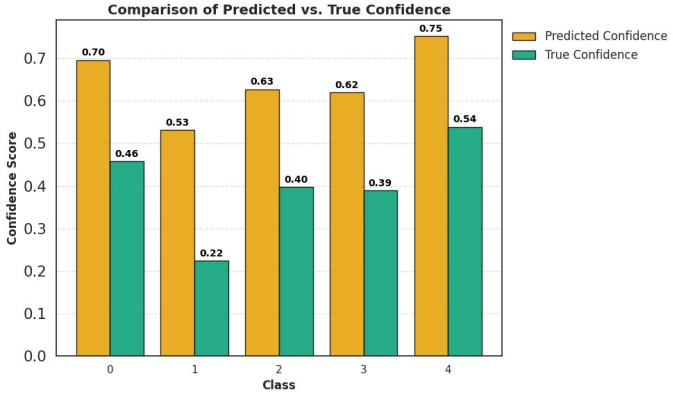

Figure 10: Comparison of predicted confidence and true confidence for every class.

The second plot (10) compares **Predicted Confidence** (orange bars) with **True Confidence** (green bars) across different classes:

- **Predicted Confidence:** For each predicted class, we compute the **average confidence score** of the model's predictions of the class on the test set.
- **True Confidence:** For each true class, we compute the **average confidence score** that the model assigned to the true class when it made the prediction.

This distinction allows us to assess whether the model is **well-calibrated**, meaning whether its confidence reflects the actual likelihood of correctness.

The largest confidence gap appears in **Class 1**, where the model's predicted confidence is far higher than its true confidence. This indicates that the model is overconfident when predicting this class, which may lead to a higher misclassification rate.

6 Conclusion

The results indicate that models trained on zoomed-in nodules consistently outperform those trained on full-slice images. Additionally, binary classification models demonstrated higher accuracy than their multi-class counterparts. This was expected,

as zoomed-in images provide a more focused view of the region of interest, reducing background noise and irrelevant anatomical structures, thus facilitating more precise predictions.

A potential application of these models in a clinical setting involves combining predictions from classifiers trained on different types of images but related to the same patient. Since zoomed-in images are explicitly designed to highlight the tumor region, leveraging their strengths alongside full-slice predictions could enhance overall diagnostic robustness. By integrating the insights from both approaches, a hybrid decision-making system could be developed to balance the strengths of each model, ultimately leading to improved classification reliability.

7 Future Development

Several improvements could enhance model performance and clinical applicability:

- **Cross-Validation and Hyperparameter Optimization:** Due to computational constraints, cross-validation was not implemented but could improve model robustness. Stratified k-fold cross-validation combined with grid search may help optimize hyperparameters and enhance generalization.
- **Multi-Modal Data Integration:** Incorporating clinical data (e.g., patient history, radiomics) could enrich the model's predictive capability beyond image-based classification.
- **Ensemble and Hybrid Models:** Future work should explore ensemble techniques or fusion strategies that effectively integrate full-slice and zoomed-in classifiers, optimizing performance while ensuring clinical applicability and reliability.

Exploring these directions could enhance accuracy, robustness, and real-world applicability, contributing to improved early detection and diagnosis of lung cancer [3].

References

[1] H.-W. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. 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] Qi Zhou, Yan Li, Jian Wang, Xiaoyan Chen, and Hong Zhang. A comprehensive review on ai-driven drug discovery: Recent advances, challenges and future perspectives. *Artificial Intelligence in the Life Sciences*, 4:100055, 2024.

[3] X. Li and Y. Fan. A 3d convolutional neural network for pulmonary nodule detection. *arXiv preprint arXiv:1904.03501*, 2019.