POLITECNICO
MILANO 1863

Lung Cancer Malignacy score
DeepBreath AI

Sofia Brunori, Alberto Eusebio, Alessandro Tognotti, Andrea Tomasella

247467, 244659, 245256, 247328

February 5, 2025

1 Introduction

Lung cancer (LC) remains a leading cause of death worldwide. Early diagnosis is critical to protect innocent human lives. **Computed tomography (CT)** scans are one of the primary imaging modalities for LC diagnosis. However, manual CT scan analysis is time-consuming and prone to errors/not accurate. Considering these shortcomings, computational methods especially machine learning and deep learning algorithms are leveraged as an alternative to accelerate the accurate detection of CT scans as cancerous or non-cancerous. The objective of this assignment is to develop classifiers for LC malignancy detection and confidence estimation using different input types. The classification task involves distinguishing between **benign** (classes 1, 2, and 3) and **malignant** cases (classes 4 and 5), as well as performing a finer 5-class malignancy classification. These tasks will be performed both using the full-slice input and the zoomed-slice input, each with an associated confidence estimate to assess the robustness of the model's predictions. Furthermore, a comparative analysis between the full-slice and zoomed-slice input will be conducted to evaluate performance variations and determine the impact of input selection on classification accuracy and confidence.

2 Materials

2.1 Dataset description

The dataset used consists of CT scan slices from **2363** patients, where each patient is represented by a full CT slice and a zoomed-in nodule slice. The full slices are all **512×512** 2D matrices of **int16** values, while the zoomed-in nodule slices have varying dimensions but are also stored as 2D matrices of int16 values. The nodule slices are cropped regions extracted from the original full-slice images, with the tumor centered in each cropped region. All the slice files were provided in the **NRRD** format as matrices of **HU values**. After an analysis of the labels distribution, it was observed that the dataset is highly **imbalanced**. In the binary classification task, where tumors are categorized as benign or malignant, the distribution of classes is 1793 benign samples and 570 malignant samples. In the multi-class classification task, where tumors are assigned a malignancy score from 1 to 5, the class distribution is as follows: 244 (class 0), 457 (class 1), 1092 (class 2), 418 (class 3), and 152 (class 4). The imbalance in the dataset influences the choice of techniques and methodologies applied in model training, as different approaches are required to handle the varying class distributions effectively.

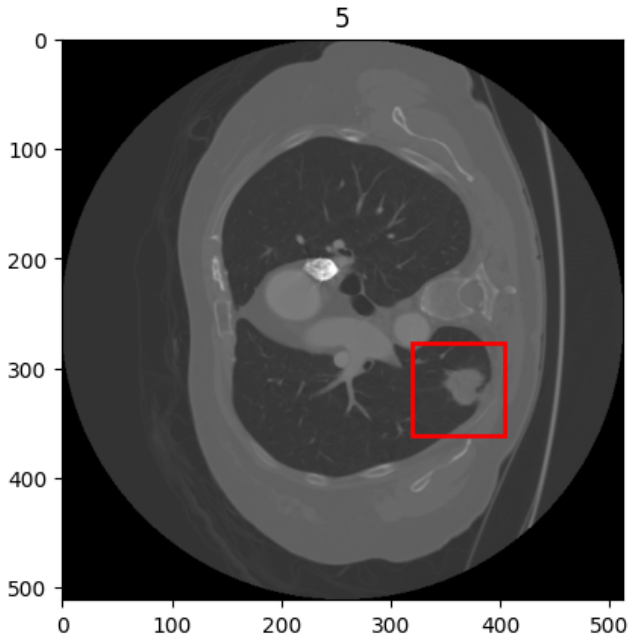


Figure 1: Full slice zoomed with patch

2.2 CT Scans and Hounsfield Units

Computed Tomography (CT) scans are medical imaging techniques that use X-ray beams to capture detailed cross-sectional images of the human body. Unlike traditional X-rays, which provide 2D projections, CT scans reconstruct 3D representations of internal structures, allowing for precise visualization of tissues, organs, and abnormalities.

A key feature of CT imaging is the use of **Hounsfield Units (HUs)**, a standardized scale that quantifies tissue density based on X-ray attenuation. The Hounsfield scale assigns specific values to different materials:

- **Air**: -1000 HU
- **Lung tissue**: -500 to -700 HU
- **Water**: 0 HU (reference point)
- **Soft tissues**: 30 to 100 HU
- **Bone**: 700 to 3000 HU

These values help differentiate between various anatomical structures and pathological conditions. For example, tumors may appear at different HU ranges depending on their composition, and contrast-enhanced scans allow for better tissue differentiation.

3 Method

3.1 Data Preprocessing

For the nodule classification task, preprocessing steps were applied to ensure consistency across images. All nodule images were resized to have a uniform shape. The dataset was then split into training, validation, and test sets, with **72%** of the total dataset allocated for training, **8%** for validation, and **20%** for testing. A min-max normalization was applied using the minimum and maximum values from the training set, resulting in images scaled within the $[0, 1]$ range. These normalized images were then multiplied by 255 to restore uint8 values format. Finally, grayscale images were converted to RGB format by stacking the same channel three times, ensuring compatibility with deep learning models requiring three-channel input.

For the full CT slice preprocessing, a **windowing transformation** was applied to enhance the visibility of lung tumors. The original HU values were transformed using specific windowing parameters tailored to each classification task.

For both the **5-class malignancy classification** and the **binary classification** (benign vs. malignant), a **window width** (W) of 1300 and a **window level** (L) of -220 were applied. These settings were chosen to enhance tumor visibility by optimizing contrast and reducing irrelevant intensity variations, ensuring better differentiation of malignancy features across both classification tasks. These values were inspired by the analysis of Yao [8]. The minimum and maximum intensity values were so computed as:

$$v_{\min} = L - \frac{W}{2}, \quad v_{\max} = L + \frac{W}{2}$$

This means that HU values outside these ranges were clipped accordingly. After clipping, the values were normalized to a $[0, 255]$ grayscale range using linear rescaling, making the images more suitable for deep learning processing. The resized grayscale images were then transformed into three-channel images by stacking the same grayscale values across three layers to ensure compatibility with deep learning architectures. This preprocessing pipeline ensured that all CT slices were standardized in terms of intensity distribution and image format, improving the robustness of the classification model. For the binary classification task, tumor classes were

converted into a binary format: 0 - benign (class 1, 2, 3) and 1 - malignant (class 4, 5). In contrast, for the multiclass classification task, the original labels were preserved.

3.2 Model architectures

The differences between datasets and classification tasks necessitated the adoption of distinct techniques and methodologies. The following section outlines the selected model architectures and the corresponding training steps for each classification approach

3.2.1 Full-slice Binary Classification

To handle class imbalance, **RandomOversampling** [3] was applied, in order to correct the class imbalances in the train set. In the data augmentation pipeline `ImageDataGenerator` was used to apply the following transformations:

```
ImageDataGenerator(
    rescale=1/255,
    rotation_range=10,
    width_shift_range=0.01,
    height_shift_range=0.01,
    zoom_range=(0.95,1),
    brightness_range=(0.95,1.05))
```

These augmentations help diversify the training data, improving the model's robustness and generalization. For the binary classification task using full-slices, a **transfer learning approach** was adopted by leveraging the **EfficientNetV2S** [7] model pretrained on the ImageNet dataset. The last **10 layers** were unfrozen, to allow fine-tuning on the dataset. After the pretrained backbone, a **Global Average Pooling** layer was incorporated to reduce feature dimensionality while retaining essential information. This was followed by a **fully connected** layer with **128** units and **ReLU** activation, regularized using **L1** (0.1) **L2** (0.01) penalties to enhance generalization. To further mitigate overfitting, a **Dropout** layer (0.5) was applied, along with **Batch Normalization** to stabilize training. Finally, the model outputs a single neuron Dense layer using a sigmoid activation. The model was trained using **Binary Focal Crossentropy** [4] loss and **Adam** optimizer, with a learning rate of 1×10^{-5}. Two callbacks were employed at training stage: **EarlyStopping** on the validation accuracy, with patience set to 25, and **ReduceLrOnPlateau** on the validation loss, with a patience of 10.

3.2.2 Full-slice multiclass classification

To address class imbalance in the dataset, **RandomOversampling** was performed to match the number of samples in each class to that of the majority class in the training set. The classification model was built upon **EfficientNetB1** [6], a pretrained convolutional neural network originally trained on ImageNet, as performed in the paper of Raza [5]. The pretrained weights were used to initialize the model, and all layers were fine-tuned on the training images.

The model was trained using a **Focal Loss** with alpha = 0.6 and gamma = 2.0, ensuring that the network focused on correctly predicting all five classes rather than being biased toward the majority class. The **Adam** optimizer was employed with a learning rate of 5×10^{-5} to facilitate stable convergence. Two callbacks were employed at training stage: **EarlyStopping** on the validation accuracy, with patience set to 8, and **ReduceLrOnPlateau** on the validation loss, with a patience of 5. As classification head, two fully connected (Dense) layers with 128 and 64 neurons, respectively, were added. These were followed by **Dropout** and **Batch Normalization** layers to mitigate overfitting. The output layer consisted of five neurons with a **softmax** activation function, enabling multi-class probability distribution prediction.

During training, data augmentation was applied using a **image data generator** with the following configuration:

```
ImageDataGenerator(
    rotation_range=10,
    width_shift_range=0.1,
    height_shift_range=0.1,
    zoom_range=(0.9,1.1),
    brightness_range=(0.9,1.1))
```

3.2.3 Zoomed-slice binary classification

In the binary classification task using zoomed slices, an imbalance in the dataset was observed, with sig-

nificantly more benign cases than malignant ones. To mitigate this issue, **Synthetic Minority Over-sampling Technique (SMOTE)** [1] was applied. Class weights were computed to assign higher weights to underrepresented classes to help balance their influence during training. SMOTE was then applied to oversample the minority class. This approach allowed for a more balanced dataset, improving the model's ability to learn features from both benign and malignant samples without being biased towards the majority class.

For the binary classification task using zoomed slices, a **transfer learning approach** was adopted by leveraging the **EfficientNetV2S** [7] model, pre-trained on ImageNet, as a feature extractor. The pre-trained model was loaded with its weights, and the last **10 layers were unfrozen** to allow fine-tuning on the dataset.

To enhance model generalization, extensive **data augmentation** was applied using `ImageDataGenerator`.

```
ImageDataGenerator(
    rescale=1/255,
    rotation_range=360,
    width_shift_range=0.2,
    height_shift_range=0.2,
    shear_range=0.2,
    zoom_range=(0.7, 1.3),
    horizontal_flip=True,
    vertical_flip=True,
    brightness_range=(0.6, 1.4),
    fill_mode='nearest',
    zoom_range=(0.95,1),
    brightness_range=(0.95,1.05))
```

After the pretrained backbone a **Global Average Pooling layer** was introduced to reduce feature dimensionality, followed by a **fully connected layer** with 128 units and **ReLU activation**. To prevent overfitting, **Dropout (0.5**) and **Batch Normalization** are applied. The final **Dense layer** has a single neuron with a sigmoid activation function for binary classification.

The model was compiled with **binary crossentropy loss**, the **Adam optimizer** (with a learning rate of 1×10^{-5}), and the metrics Accuracy and AUC.

3.2.4 Zoomed-slice multiclass classification

In the multiclass classification task using zoomed nodule crops, a strong class imbalance was observed . After testing various balancing techniques, including class weights, undersampling, and oversampling, the most effective approach combined **SMOTE and data augmentation**.

SMOTE was applied to all classes, adding **250 synthetic samples** per class to enhance diversity and increase train set size. Then, **data augmentation** was performed to adjust all classes to match class 3's sample count using geometric transformations ($\pm 30°$ rotations, $\pm 10\%$ translations, flipping). To prevent overfitting, one-third of class 3's samples were replaced with augmented versions, ensuring greater variability and improved generalization.

During training, **ImageDataGenerator** dynamically applied extensive augmentation to training batches, following the same strategy as in the binary classification task. This approach prevented overfitting and improved the model's ability to generalize to unseen data.

For this task, a transfer learning approach was employed by utilizing the **EfficientNetB0 model** [6], pre-trained on ImageNet, as a feature extractor. The pre-trained model was initialized with its pre-trained weights, and the last **20 layers** were unfrozen to enable fine-tuning on the dataset. After the pretrained backbone a **Global Average Pooling layer** was introduced to reduce feature dimensionality, followed by **Dropout (0.2)** and **Batch Normalization** to prevent overfitting. The architecture also contains a **fully connected layer** with 64 units and **ReLU activation**, another **Dropout (0.2) layer**, and the final **Dense layer** with 5 neurons and **softmax activation** for multiclass classification.

The model was compiled using **sparse categorical crossentropy** as the loss function, and the **Adam optimizer** set to a learning rate of 1×10^{-3}. Two callbacks were employed at training stage: **EarlyStopping** on the validation accuracy, with patience set to 30, and **ReduceLrOnPlateau** on the validation loss, with a patience of 8. These strategies helped stabilize training, improving convergence while avoiding overfitting and excessive training time.

Table 1: The results obtained from the different models and tasks

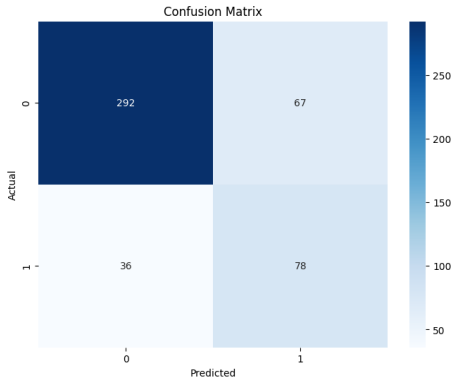
Task	Accuracy	Precision	Recall	F1	AUC
Full binary	0.73 ±0.02	0.44 ±0.03	0.52 ±0.05	0.48 ±0.03	0.66 ±0.02
Full multiclass	0.23 ±0.04	0.20 ±0.04	0.21 ±0.05	0.19 ±0.04	0.51 ±0.04
Zoomed binary	0.72 ±0.03	0.45 ±0.05	0.63 ±0.08	0.52 ±0.05	0.75 ±0.04
Zoomed multiclass	0.53 ±0.02	0.46 ±0.04	0.44 ±0.03	0.43 ±0.03	0.74 ±0.02

4 Results

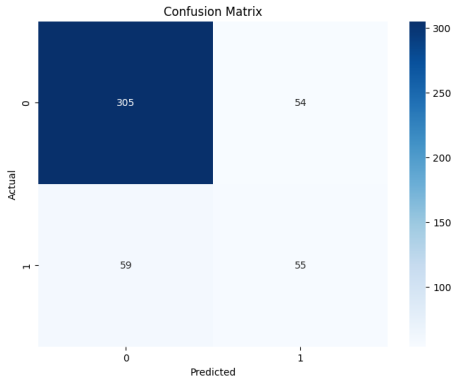
The models were all trained using the previously described train, validation and test split. To assess their performance, the Monte Carlo Dropout method [2] was employed. Initially, the models were assessed on the validation split to determine the best-performing model for each task. Subsequently, the models were evaluated on the test split **50** times, leaving the last Dropout layer of the architecture active. The results of each run were averaged and the 95% confidence intervals were obtained for each metric. The results are reported in Table1.
In the Table1 are reported the metrics of accuracy, precision, recall, F1 score and the AUC score. Among these, the AUC score was selected as the most representative metric due to its robustness in the context of imbalanced datasets. In fact, while accuracy can sometimes be misleading when the data is highly skewed towards one class, the AUC score provides a more balanced evaluation by considering both the true positive rate (sensitivity) and the false positive rate.

As illustrated in the figures the confusion matrices clearly show that the model is highly influenced by class imbalance, as it tends to bias its predictions towards the more frequent class, despite efforts to balance the dataset. This trend is observed both in the binary classification task and, more significantly, in the multi-class task.

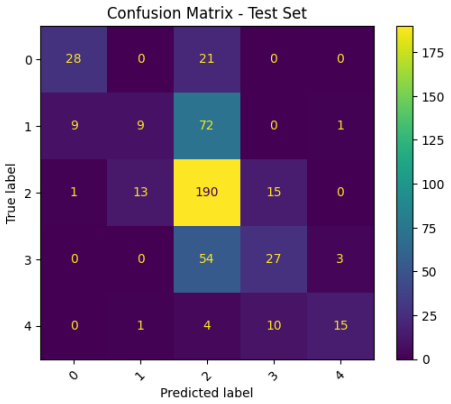
As shown in the Table1, the AUC values vary across different tasks and input configurations. The varying AUC can be attributed to several key factors. The model's ability to differentiate between classes is influenced by the level of detail in the input data. For the multi-class task with full slices, the model faces greater difficulty due to the complexity of distinguishing between multiple classes, particularly when the dataset is imbalanced. On the other hand, focusing on zoomed-in regions of the images for both binary and multi-class tasks improves the model's ability to detect relevant features. By narrowing the input to smaller areas, the model is able to focus more effectively on discriminative patterns, which enhances its AUC, especially in binary classification. However, when the model is required to perform multi-class classification on zoomed-in images, the complexity increases, and the AUC naturally drops. Additionally, the class distribution and the dataset's characteristics likely play a role in the observed AUC trends, with more complex tasks being affected by the model's ability to generalize across multiple categories.



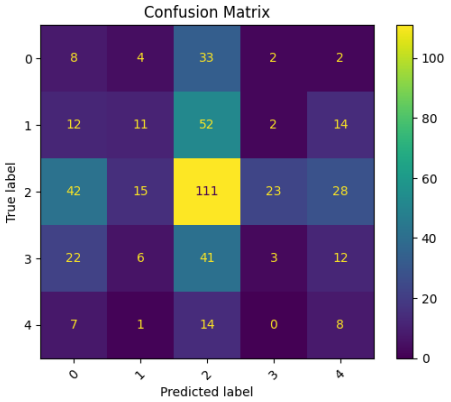
(a) Confusion Zoom Binary



(b) Confusion Full Binary



(a) Confusion Zoomed Multiclass



(b) Confusion Full Multiclass

5 Conclusions

The results clearly show that the best-performing models were on the binary classification tasks. The high imbalance in the dataset made the training insidious, and the models clearly struggled in the classification of the minority classes. Using class weights helped the model avoid overfitting on a single prediction and improved generalization. Consistently with expectations, the best results were obtained on the zoomed-slices binary classification task, and the worst was found on the full-slice multiclassification tasks, as they were respectively the most simple and most difficult tasks. In fact, in the full-slice classification task, the model had to learn to localize the tumor and classify it. It is also worth noticing that zoomed-slice classification could benefit more on data augmentation with respect to full-slice, as in the latter case, a stronger augmentation could easily make the tumor invisible or largely deformed.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[2] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

[3] M. L. C. Lauron and J. P. Pabico. Improved sampling techniques for learning an imbalanced data set, 2016.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018.

[5] R. Raza. Lung-effnet: Lung cancer classification using efficientnet from ct-scan images. *Engineering Applications of Artificial Intelligence*, (126), 2023.

[6] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[7] M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[8] G. Yao. Value of window technique in diagnosis of the ground glass opacities in patients with non-small cell pulmonary cancer. *Oncology Letters*, 12(5):3933–3935, 2016.