POLITECNICO
MILANO 1863

Applied AI in Biomedecine
Medica Nova

Team members: Bechet de la Peschardière Aurore, Noël Malvin, Gallego Toscano Olivia

Matricola codes: 276600, 276847, 276398

January 10, 2025

1 Introduction

This paper serves as the final report for the course Applied AI to Biomedicine at the Politecnico di Milano for the academic year 2024/2025. The project focused on developing deep learning models to classify CT scan images by malignancy scores, employing advanced techniques to improve the accuracy of lung cancer diagnosis in unseen data. We are provided of two images per patient : a full-slice one and a zoomed one which we will refer to as the nodule image. The process began with data set preparation, followed by data pre-processing and a split between train-validation-test. Feature extraction was then performed to enhance the representativeness of the dataset. Subsequently, we implemented a range of models, starting with baseline models as a foundational benchmark, followed by advanced models, and finally deep learning architectures. A comparative analysis was performed to determine the model that achieved the highest precision and precision.

2 Materials

We began by thoroughly examining the provided data set. This data set contains approximately 2300 patients, each patient associated with two CT images: a full slice and a zoomed slice that focuses on the largest nodule area. In addition, malignancy scores (ranging from 1 to 5) are provided in an accompanying Excel file.

2.1 Data preparation

As an initial step, we loaded and verified the consistency of the dataset, ensuring that each patient had exactly one full-slice image and one zoomed-slice image, as well as a corresponding malignancy score. To better understand the dataset, we visualized the label distribution and plotted example images from each class.

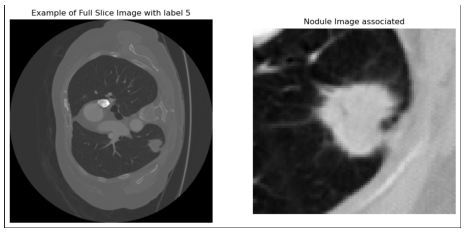

Figure 1: Example of data

We also checked the validity of the labels, confirming that all malignancy scores fell within the expected range (1 to 5).

An important observation was the dataset imbalance, with a significantly higher number of benign cases (scores 1-3) compared to malignant cases (scores 4-5). This imbalance highlights the need for appropriate strategies during the training phase to ensure fair learning.

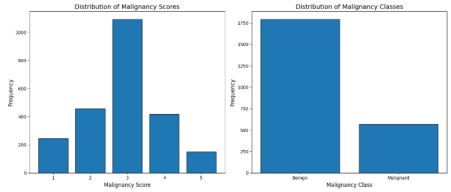

Figure 2: Distribution of Malignancy scores and classes

To prepare the data for classification, a new column indicating malignancy was created, categorizing nodules as either benign or malignant based on their malignancy score on the 1-to-5 scale. Additionally, all images were resized to a uniform size to ensure consistency across the

dataset, facilitating efficient processing by machine learning models. Pixel intensity values were also normalized to a range between 0 and 1, further standardizing the data. These preprocessing steps ensured a uniform and robust dataset for accurate model training and evaluation.

Since the dataset was imbalanced, with more benign cases than malignant ones, we computed class weights to counteract this imbalance during training.

2.2 Feature extraction

To fully exploit the images of the dataset, we extracted advanced features from the full slices and nodules. These features included:

- **First-order statistics** : mean intensity, variance, skewness, kurtosis, minimum, maximum, entropy of pixel values
- **Textural features using the gray-level co-occurrence matrix (GLCM)** : contrast, dissimilarity, homogeneity, angular second moment (ASM), energy, correlation.
- **Wavelet-based features capturing localized frequency information** : calculate approximation and detail coefficients, capturing frequency-domain characteristics.
- **Morphological features (only for nodule features)**: Properties such as area, perimeter, equivalent diameter, circularity, irregularity, elongation, convexity, and solidity. These features were calculated based on a segmentation of the nodule regions, which allowed us to isolate the nodules and focus specifically on their morphology.

Figure 3: Segmentation of nodules

Once the features are computed they are stored in a dataframe which will be used in our models.

3 Methods

Our methodology consisted of several key stages that were designed to optimize the performance of the model

3.1 Splitting of the data

Before testing any models we split the data to have a training set (64%), a validation set (16%) and a test set (20%), while preserving the label distribution using stratified sampling.

3.2 Classification

We used four classical machine learning models to evaluate their predictive performance on our data. We began with baseline models, which are simpler: Logistic Regression and Decision Tree. Subsequently, we implemented more advanced models like Random Forest.

First of all, to address the class imbalance in the dataset, class weights were computed and applied to all models to ensure fair representation of minority classes. The calculated weights were: {`4: 5.66, 3: 2.16, 1: 9.69, 2: 5.16, 5: 15.49`}.

3.2.1 Full Slice - 5 class classification

We started our development by focusing on the full slices images.Our aim is to develop a classification model which output will be the 5 class malignancy score. Once this classification model is accurate we will only need to add one more layer which will take the malignancy score predicted by the network and return the binary result (benign or malignant) associated. The performance metrics of these models are detailed in Table 1. The following subsections summarize the observations for each model:

1. **Logistic Regression**: The model achieved an accuracy of 15.2%. It showed the highest performance for class 3, with a precision of 0.45 and a really low recall of 0.009, while struggling with class 5, with a precision of 0.07 and a recall of 0.33.
2. **Decision Tree**: The model gave an overall accuracy improvement, getting a 29.4%, but it was still low. As well, it showed an extremely poor performance for class 5 and low F1-scores for classes 1, 2, and 4.
3. **Random Forest (RF)**:For this model, we tried different versions:
   (a) **Basic RF**: This model achieved an overall accuracy of 45.67% and an F1-score of 25%. It performed particularly well on class 3, with a recall of 0.83 and an F1-score of 0.61. However, it struggled with classes 2, 4, and 5, where both precision and recall were very low.
   (b) **RF cross-validated with the Bayesian Theorem**: This version gave us an slightly lower accuracy that the previous one, 45.03%.

Even thought the performances improved for class 1 and maintained for class 3.

(c) **RF cross-validated with Grid Search**: This model achieved an overall accuracy of 44% and a macro F1-score of 26%. It performed particularly well on class 3, with a recall of 0.80 and an F1-score of 0.59. However, it struggled significantly with class 5, where both precision and recall were 0. Similarly, classes 1, 2, and 4 showed poor performance, with low recall values, highlighting the model's difficulty in distinguishing these classes effectively.

(d) **RF cross-validated with Optuna + Probability Calibration**: Using Optuna, the Random Forest model underwent 100 trials for hyperparameter optimization, aiming to maximize the macro F1-score. With these parameters, the model achieved a best macro F1-score of 0.2819 during cross-validation.

Following optimization, Platt Scaling was applied for probability calibration to improve the reliability of probability estimates. The calibration metrics for each class (Classes 1–5) were as follows:

- **Class 1**: Brier Score Loss = 0.0874, ROC AUC = 0.7066
- **Class 2**: Brier Score Loss = 0.1535, ROC AUC = 0.5825
- **Class 3**: Brier Score Loss = 0.2447, ROC AUC = 0.5553
- **Class 4**: Brier Score Loss = 0.1375, ROC AUC = 0.6715
- **Class 5**: Brier Score Loss = 0.0598, ROC AUC = 0.7332

These results indicate that Class 5 had the best-calibrated probabilities with a low Brier Score Loss and high ROC AUC, while Class 3 showed poorer calibration performance.

Relevance of Optuna and Probability Calibration: Optuna's efficient hyperparameter optimization enabled the identification of a near-optimal Random Forest configuration with minimal computational cost. Probability calibration added a layer of reliability, ensuring that the model's predicted probabilities aligned more closely with actual probabilities. This is especially critical for lung tumor prediction, where calibrated probabilities provide interpretable and trustworthy confidence estimates, enhancing clinical decision-making.

Table 1: Performance Metrics of Classical Machine Learning Models.

Model	Accuracy	Precision (avg)	Recall (avg)	F1 (avg)
Logistic Regression	0.15	0.29	0.15	0.16
Decision Tree	0.29	0.28	0.29	0.29
RF Basic	0.46	0.39	0.46	0.38
RF+Bayesian	0.45	0.40	0.45	0.39
RF+GridSearch	0.44	0.38	0.44	0.37

3.2.2 Nodule - 5 class classification

After analyzing the results of the full slice classification, we shifted our attention to the study of nodules, aiming to refine the model further. Nodules represent a more localized and detailed region of interest, which could offer critical insights and potentially improve the malignancy classification performance. The following subsections summarize the methodology and observations derived from the nodule-based study:

- **Logistic Regression**: The model achieved overall accuracy of 34%. It showed the best performance for class 3, with a precision of 0.58 and a recall of 0.34. However, the resutls for the remaining classes were poor.
- **Decision Tree**: The model achieved an accuracy of 33%. It performed best with class 3, showing a precision of 0.48 and a recall of 0.43, but struggled with class 2, with a precision of 0.18 and a recall of 0.16.
- **Support Vector Machine (SVM)**:
- **Random Forest (RF)**:For this model, we tried different versions:

(a) **Basic RF**: This model achieved an overall accuracy of 54% and an F1-score of 47%. It performed particularly well on class 3, with a recall of 0.92, a high precision of 0.56, resulting in a strong F1-score of 0.70. However, it struggled with classes 2, and the performance for the remaining classes was moderate.

(b) **RF cross-validated with the Bayesian Theorem**: This version achieved a slightly lower prediction accuracy compared to the previous one, at 44%, indicating that the hyperparameter tuning did not significantly enhance the model's performance. The best-performing class remained Class 3, with

a precision of 0.53, recall of 0.69, and an F1-score of 0.60. On the other hand, the model performed poorly on Class 2 and showed only moderate performance on the remaining classes.

(c) **RF cross-validated with Grid Search**: This model achieved an overall accuracy of 56% and a macro F1-score of 45%. It showed strong performance on class 3, with a recall of 0.85 and an F1-score of 0.68. Class 5 also performed relatively well, with a recall of 0.47 and an F1-score of 0.50. However, the model faced difficulties with class 2, which had a low recall of 0.09 and an F1-score of 0.15, indicating difficulties in effectively identifying this class. Classes 1 and 4 had moderate performance, reflecting the model's overall balanced yet suboptimal predictive capabilities.

(d) **RF Cross-Validation with Optuna and Probability Calibration**: The Random Forest model was optimized using Optuna to maximize the macro F1-score. After hyperparameter tuning, the model was retrained on the full training dataset, followed by probability calibration to evaluate the reliability of its predictions.
During cross-validation, the highest macro F1-score achieved was 0.4488.
After retraining, probability calibration metrics were computed for each class:

(e) **Class 1**: Brier Score Loss = 0.0670, ROC AUC = 0.8739

(f) **Class 2**: Brier Score Loss = 0.1493, ROC AUC = 0.6449

(g) **Class 3**: Brier Score Loss = 0.2143, ROC AUC = 0.7118

(h) **Class 4**: Brier Score Loss = 0.1191, ROC AUC = 0.8164

(i) **Class 5**: Brier Score Loss = 0.0409, ROC AUC = 0.9333 These results indicate that

the model produced the most reliable predictions for Class 5, with the lowest Brier Score Loss and the highest ROC AUC. Conversely, Class 3 exhibited the weakest calibration performance. The Optuna-tuned Random Forest model underscored

Table 2: Performance Metrics for Nodule Classification Models.

Model	Accuracy	Precision (avg)	Recall (avg)	F1(avg)
Logistic Regression	0.34	0.41	0.34	0.35
Decision Tree	0.33	0.35	0.33	0.34
SVM	0.00	0.00	0.00	0.00
RF Basic	0.46	0.47	0.54	0.47
RF+Bayesian	0.44	0.40	0.44	0.41
RF+GridSearch	0.56	0.54	0.56	0.51

3.3 Deep Learning Methods

3.3.1 Preprocess data

Before implementing different deep learning models, we made adjustments to the data splitting and preprocessing. First, we transitioned from using tabular features in Excel to working directly with image data, adding a channel dimension to the images to represent the grayscale color range. Next, we revised the data split into training (60%), validation (20%), and test (20%) sets, ensuring that class distribution (stratification) was maintained across all subsets. Additionally, we standardized the label range from 0 to 4, as TensorFlow requires labels to start from 0 for multi-class classification tasks.

3.3.2 CNN Model

The CNN achieved an accuracy of 46%, as shown in the graphs of Figure 5. From the plots of training and validation accuracy and loss, we observe significant fluctuations in accuracy and high volatility in loss values, which could indicate issues such as overfitting, underfitting, or unstable training dynamics. Based on these results, we decided to explore other models like EfficientNetB3, ResNet, and VGG, which may capture the spatial and textural features of the dataset more effectively and potentially improve overall performance.

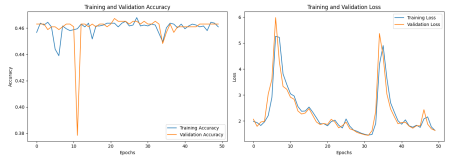

Figure 5: CNN Training and Validation analysis

3.3.3 EfficientNetB3 Model

The EfficientNetB3 model, achieved a test accuracy of 38% and a weighted F1-score of 0.36, this showed that the model struggles to generalize effectively across all classes. This is evident from the low recall and precision scores for most classes, indicating an imbalance in its ability to accurately classify minority classes. Given these limitations, we will transition to trying the EfficientNetV2L model, which offers enhanced architecture and improved feature extraction capabilities. The aim is to leverage its more advanced design to better capture complex patterns in the dataset and improve overall classification performance.

3.3.4 EfficientNetV2L Model

The EfficientNetV2L model demonstrated poor performance, with a test accuracy of only 15% and a weighted F1-score of 0.08. The model failed to generalize across most classes, as reflected in the low recall and F1-scores for classes 2 and 3. While class 1 showed a recall of 0.47, its precision and F1-score remained low, indicating significant misclassification issues. Similarly, classes 4 and 5 exhibited moderate recall but poor precision, resulting in overall weak performance. These results suggest that the model struggled to extract meaningful features from the data, necessitating further improvements in architecture or preprocessing to achieve better classification results. Because of this, we decided to try ResNet50, hoping to take advantage of its strengths—such as powerful feature extraction, residual learning, and better generalization—to overcome the limitations of the previously tested models and boost overall classification performance.

3.3.5 ResNet50 Model

The ResNet50 model achieved a test accuracy of 46% and a weighted F1-score of 0.29. While it performed exceptionally well for Class 3, with a perfect recall (1.00) and an F1-score of 0.63, it completely failed to classify the other classes, showing zero precision, recall, and F1-scores for Classes 1, 2, 4, and 5. This suggests that ResNet50 is heavily biased toward Class 3, likely due to class imbalance, and struggles to generalize across the remaining classes. As a result, the model may perform adequately when Class 3 is predominant, but it lacks reliability for accurately identifying other tumor types. To address this issue, we moved to the VGG16 model, aiming for more balanced feature learning and better classification of minority classes through its simpler architecture and potential for improvement.

3.3.6 VGG16 Model

The VGG16 model achieved a test accuracy of 45% and a weighted F1-score of 0.33. While it performed well for Class 3 with a recall of 0.93 and an F1-score of 0.62, the model struggled significantly with other classes, showing near-zero recall and F1-scores for Classes 4 and 5. This highlights a strong bias toward the majority class, indicating that further adjustments, such as addressing class imbalance, are needed to improve performance across all classes. To tackle this, we shifted to binary classification to simplify the problem and improve model performance.

Table 3: Performance Metrics of Deep Learning Models.

Model	Accuracy	Precision (avg)	Recall (avg)	F1(avg)
EfficientNetB3	0.28	0.37	0.38	0.36
EfficientNetV2L	0.15	0.54	0.15	0.08
ResNet50	0.46	0.21	0.46	0.29
VGG16	0.45	0.31	0.45	0.33

4 Full Slice - Binary Classification

We selected Random Forest, the best-performing model from our previous experiments, to handle the binary classification task. Using Optuna, we optimized its hyperparameters and trained the model on the full dataset. The model achieved a best F1-score of 0.7098, and its reliability was further evaluated through calibration curves.

Table 4: Calibration Metrics for Full Slice - Binary Classification.

Metrics	Value
Brier Score Loss	0.1765 (lower is better)
ROC AUC Score	0.6180

Figure 6 illustrates the calibration curve, showing the relationship between predicted and true probabilities. The green dashed line represents perfect calibration, the blue line shows uncalibrated predictions, and the orange line represents calibrated probabilities. The calibrated model aligns more closely with the perfect calibration line, enhancing prediction reliability.

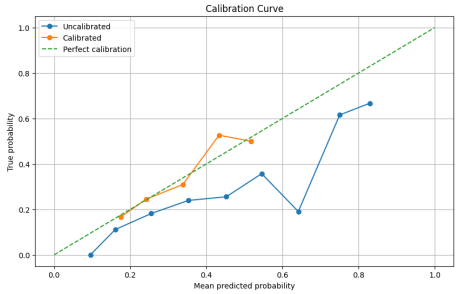

Figure 6: Calibration Curve for Full Slice - Binary Classification.

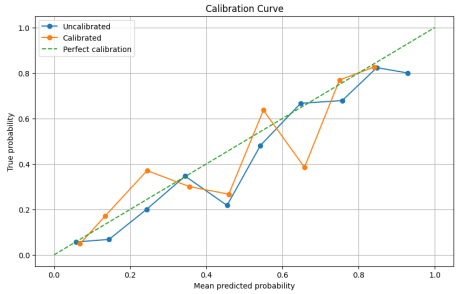

Figure 7: Calibration Curve for Nodule - Binary Classification.

5 Nodule - Binary Classification

The nodule-based binary classification task focuses on distinguishing benign nodules (malignancy scores 1–3) from malignant ones (malignancy scores 4–5). Again, the Random Forest model was chosen for its strong performance. Hyperparameters were optimized with Optuna to maximize the macro F1-score. The final model was retrained on the full training set, evaluated on the test set, and calibrated to improve the reliability of its predictions.

Best F1-score: 0.8401

Table 5: Calibration Metrics for Nodule - Binary Classification.

Metrics	Value
Brier Score Loss	0.1256 (lower is better)
ROC AUC Score	0.8410

Table 5 shows the model's strong performance, with a low Brier Score Loss and high ROC AUC Score, indicating accurate and reliable probability predictions.

Figure 7 shows the calibration curve, where the green dashed line represents perfect calibration. The orange line (calibrated model) is significantly closer to the green line compared to the blue line (uncalibrated predictions), confirming improved reliability after applying Platt Scaling.

The high F1-score and enhanced calibration demonstrate the model's suitability for clinical applications, where accurate and interpretable predictions are essential. This improvement is attributed to the localized focus on nodules enabled by segmentation, which provides more relevant features compared to full-slice images.

6 Discussion

The results of this study highlight the significance of both classical machine learning and deep learning models in medical image classification. While deep learning offers state-of-the-art techniques, classical models like Random Forest, when combined with feature engineering and probability calibration, can deliver competitive performance, especially with imbalanced and small datasets.

6.1 Key Observations

- **Classical Models:** Random Forest was the most effective classical model, particularly with Optuna-based hyperparameter optimization and calibration. Logistic Regression provided some baseline differentiation of classes but was limited in complexity.
- **Deep Learning Models:** EfficientNet and ResNet underperformed compared to classical models, constrained by dataset size and class imbalance. These models require larger datasets and better balancing strategies.
- **Feature Engineering:** Morphological and textural features enhanced classification, especially in nodule-based models, providing better interpretability and discriminative power.
- **Calibration and Confidence:** Probability calibration improved prediction reliability,

critical in clinical settings. Confidence estimates from calibrated models help practitioners make informed decisions.

- **Dataset Challenges:** Class imbalance posed hurdles. While class weighting and stratified sampling alleviated some issues, advanced resampling techniques like SMOTE could improve performance.

6.2 Comparison Between Full Slice and Nodule Approaches

- **Performance:** Nodule-based models consistently outperformed full-slice models, achieving an F1-score of 0.8401 compared to 0.7098. Localized analysis provided more relevant features.
- **Complexity and Clinical Relevance:** Full-slice models were more complex and prone to overfitting due to extraneous information, while nodule-based models aligned better with clinical workflows by focusing on specific areas of interest.
- **Calibration and Interpretability:** Nodule-based models had better-calibrated probabilities, offering more reliable predictions crucial for clinical adoption.

6.3 Implications for Clinical Applications

Nodule-based models demonstrated strong potential for clinical use due to their high accuracy, interpretability, and reliable confidence estimates. By focusing on localized regions, these models align with radiological practices and improve diagnostic precision.

6.4 Limitations and Future Work

- Dataset size limited deep learning performance. Expanding datasets and incorporating synthetic data generation could address this.
- Class imbalance remains a challenge, warranting advanced resampling or augmentation techniques.
- Future work should focus on explainability methods like Grad-CAM and real-world validation to ensure clinical relevance.
- Incorporating multimodal data, such as patient history and genetics, could further enhance model performance.

7 Conclusion

This study evaluated machine learning and deep learning approaches for lung cancer classification using CT images. Random Forest, with feature engineering and probability calibration, outperformed deep learning models, particularly for nodule-based analysis.

Nodule-based models proved superior to full-slice models, offering better diagnostic accuracy, interpretability, and clinical alignment. Calibration improved confidence estimation, critical for clinical decision-making. However, challenges like dataset size and class imbalance highlight the need for future improvements, including data expansion and advanced balancing techniques.

By combining classical and deep learning methods, this project provides a good framework for developing reliable tools for lung cancer diagnosis and similar medical imaging tasks.