POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Project report

Efficient and Accurate Diagnosis of Lung Tumors in Both a Binary and Multi-Class Manner

Applied AI and Biomedicine

Authors: Giulia Elizabeth de Sanctis, Nastaran Ghaffari Elkhechi, Gianluca Villa

Academic year: 2024-2025

---

1. Introduction

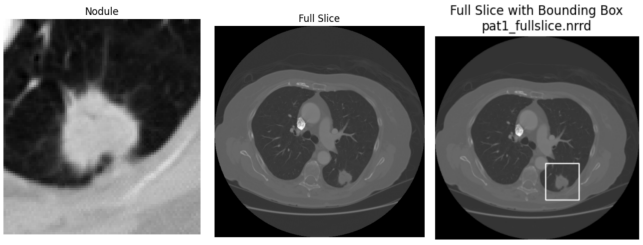
In this report, we present our approach to a classification problem aimed at diagnosing lung cancer using medical images. Lung cancer is the leading cause of cancer-related deaths worldwide, accounting for 18% of total cancer fatalities in 2020. Due to its frequent late-stage diagnosis, survival rates remain low. To address this, screening programs have been developed, increasing the demand for accurate malignant nodule detection.

Lung tumors are typically assigned a severity score from 1 to 5, with higher scores indicating greater severity. Tumors classified as 1, 2, or 3 are considered "benign," while those in classes 4 and 5 are deemed "malignant." The goal of this binary classification task is to accurately distinguish between benign and malignant tumors based on lung images.

It is important to keep in mind the medical scope of the problem. For the binary scenario, a false negative is much more serious than a false positive. A patient with a benign tumor misdiagnosed as a malignant tumor will get a big scare, but this will have no long-term effects. On the other hand, if a patient with a malignant tumor is misdiagnosed as benign, this can have dire consequences. A similar situation is presented with the five-class problem. It is much more serious and dangerous to misdiagnose a patient with a less severe tumor than the one they have as opposed to the other way around.

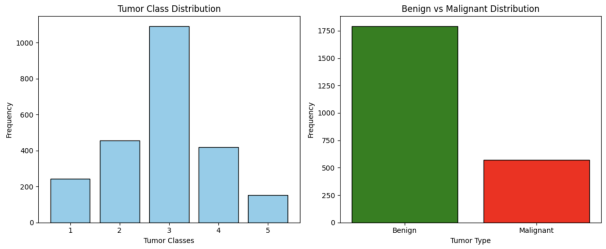
2. Dataset

Our data takes the form of images from CT scans. We are provided with two sets of images. The first is a zoom of the CT scan full on the area of interest ie the nodule. The second is an image of the full CT scan. The third image we have shown here shows the location of the nodule, ie the part of the image that the zoom concentrates on. Already from these images, the difficulty of the problem of identifying the nodule in the entire scan can be appreciated.



2.1. Exploratory Data Analysis

One of the first things we checked was the composition of our dataset as this is an important thing to consider in a problem of this type. The dataset was extremely imbalanced. The biggest class (tumors of grade 3) was about 5 times bigger than the smallest class (tumors of grade 5) and due to this, the number of images of benign tumors far outweighs the number of images of malignant tumors. This imbalance in the classes proved to be challenging.



2.2. Data Augmentation

As mentioned, the dataset is extremely imbalanced. We therefore turned to image augmentation as a way to mitigate this problem. Both the nodule and full-slice images were augmented in the same way. To begin with, all the images were normalised to be in range $[0, 1]$. A train-test-val split (80% - 10% - 10%) was then performed, since only the training set needed to be augmented. Due to the imbalance in the dataset, it was sensible to augment some classes more than others in order to achieve the goal of a more balanced dataset. Four different combinations of augmentations were selected and a subset of these was applied to each class. Augmentation 1 consisted of a vertical random flip and a zoom in. Augmentation 2 consisted of a 90 degree rotation counter-clockwise and the addition of a small constant value to every pixel in the image to make it overall darker. Augmentation 3 consisted of a horizontal random flip and a zoom out. Augmentation 4 consisted of a 90 degree rotation clockwise and the addition of a small constant value to every pixel in the image to make it overall lighter. In the five-class problem, a subset of these augmentations was applied to each class except class 3 which was the most populous. Instead, for the binary classification problem, the augmentation was only applied to the minority class, ie the CT scans of malignant tumors.

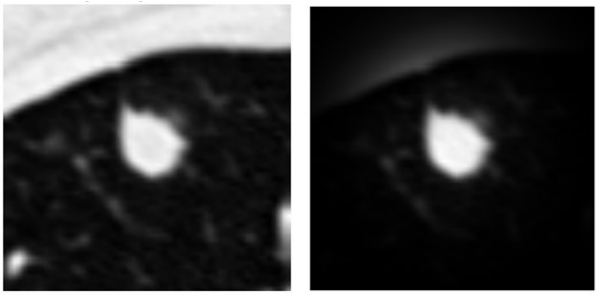
3. Nodule Classification Problem

3.1. 2-class classification problem: benign or malignant

For the Nodule Classification Problem, we used two pretrained models, InceptionResNetV2 and ResNet201, to compensate for the limited availability of data. The outputs of these models were passed through a Global Average Pooling layer, processed by Dense layers, and subsequently concatenated to form a single feature vector. This vector was then processed by a classifier. The resulting model is highly lightweight, with only 56k trainable parameters.

Given the limited amount of data available, we faced the risk of overfitting. To mitigate this, we applied L2 regularization, kernel regularization, and dropout.

Before training the model, we applied a Gaussian filter to the input images. The goal was to encourage the model to focus more on the central region of the image, where the nodule is most likely located. Below, we illustrate the effect of the Gaussian filter:



(a) Without filtering (b) With filtering

Figure 1: As we can see, the filter effectively isolates the nodule from the rest of the image, preventing the model from focusing on less relevant regions for classification, such as the peripheral areas.

The model performed remarkably well, achieving an accuracy of $86 \pm 2.2\%$ and a recall of $67 \pm 6.7\%$ on the test set. These scores were computed using bootstrap resampling on the test set observations.

We also attempted fine-tuning the model. However, this led to catastrophic forgetting where the pretrained models lost previously learned features while adapting to our dataset. This occurred because the images used for our problem are significantly different from those in ImageNet, the dataset on which the models were originally trained. As a result, during fine-tuning, the network's parameters were extensively updated to fit the new data, inadvertently overwriting the learned representations instead of integrating the new information while preserving the old knowledge.

3.2. 5-class classification problem: scale of 1-5

The 5-class model has a structure similar to the 2-class model. We used the same pretrained models, as we are dealing with a very similar task, and these models performed well in the binary classification case. To accommodate the increased number of classes, we added parameters to the classification layers and modified the output layer to account for the five classes. The performance on the test set is $53 \pm 3.2\%$.

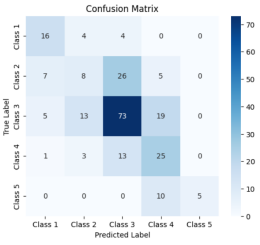


Figure 2: Confusion Matrix for the 5-Class Nodule Model

As expected, most classification errors on the test set occurred in the intermediate classes (particularly classes 2 and 3), likely because they are the most numerous in the dataset.

4. Full slice Classification Problem

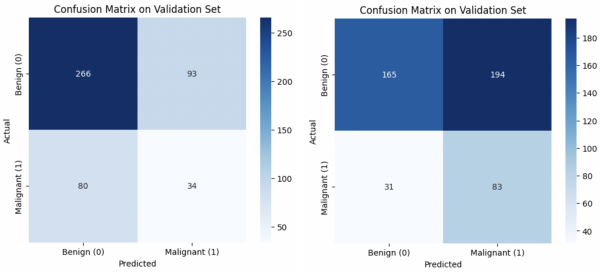
The full slice classification problem proved to be more challenging than the nodule classification problem, because more information is present in the image and the model may focus on parts that are not as relevant or may accidentally identify unhelpful patterns. We tried many different techniques to mitigate these effects.

4.1. 2-class classification problem: benign or malignant

As we mentioned before, pretrained models are used for our base models. The model seleected here was the pre-trained EfficientNet B0. This model was chosen because compared to other models in the EfficientNet family it is lightweight and thus suitable for a resource-constrained problem like ours. The EfficientNetB0 architecture was used without its classification head to extract meaningful features from the images. A global average pooling layer was applied to reduce the feature map into a vector of 1280 features. The dataset was divided into training and validation sets. The training data was composed of 1890 images. The validation data was composed of 473 images.

Model training was conducted using a multi-step process to optimize the performance and address class imbalance. SMOTE (Synthetic Minority Oversampling Technique) was applied to the training set to mitigate such class imbalance. Since traditional SMOTE is not directly applicable to images, instead of applying it on raw images, we applied it on the extracted feature vectors, thus obtained new synthetic feature vectors. After applying SMOTE, the training set consisted of 2,868 images.

We proceeded iteratively, complicating the model as we went. The first model made use of an XGBoost (Extreme Gradient Boosting) classifier (Figure 1.a). The second approach involved assigning higher class weights to the minority class by testing different ratios (2, 3, 4, and 5) using a 1:N scheme, where 1 represents the majority class weight and N the minority class weight. This N value was set as the `scale_pos_weight` parameter in the XGBoost classifier. Among the tested values, 5 yielded the best classification results, as illustrated in Figure 1.b. Precision, Recall, and F1-score for malignant cases improved reasonably in the second approach, F1-score going from 75% and 28% to 59% and 42% for benign and malignant classes respectively. Importantly, Recall for malignant cases increased from 30% to 73%, which is crucial since the detection of malignant cases is often more important in medical applications. The scores mentioned for benign cases also improved slightly, showing that this weighting did not negatively affect this class. The overall Accuracy increased slightly (51% → 52%) in the second approach, indicating better overall classification. However, from our point of view, Accuracy is not a good indicator for datasets with class imbalance because it fails to reflect the model's true performance on the minority class.



(a) XGBoost (b) Weighted XGB

Figure 3: Comparison between XGBoost and Weighted XGBoost

In the third attempt, we moved towards a feature-engineering approach, selecting the k most important features according to XGBoost. As a tree-based classifier, XGBoost computes feature importance using **gain**, measuring the average impurity reduction each time a feature is used for splitting, indicating its value in decision-making. Later, we compared the performance of the model by choosing the 50 most important and 100 most important features.

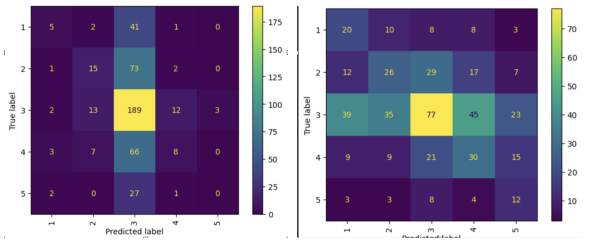
Finally, the fourth model concentrated on feature engineering using a PCA methodology. Two different models were tested: one with the features that explained 95% of cumulative variance (2 components were retained) and one with the features that explained 99% of cumulative variance (6 components were retained).

A note worthwile to be mentioned is that as expectedly, the indicated feature selection methods didn't bring us to the optimal results because selecting features solely based on importance scores or cumula-

tive explained variance as in the PCA, can lead to suboptimal classification because they ignores feature redundancy (multicollinearity) non-linear interactions, and case-specific features. Highly ranked features may be correlated, making some removals disruptive, while seemingly weak features may be crucial when combined with others.

4.2. 5-class classification problem: scale of 1-5

Once again, EfficientNetB0 was chosen as the base model for this classification, since it is a model that achieves high accuracy whilst being computationally lightweight. As an evaluation metric, the macro F1 score (the F1 score is taken for each class and averaged with the aim of balancing performance measurement) was chosen. As previously mentioned, we have a serious problem of data imbalance. One of the techniques we chose to try and compensate this is called "ADASYN" [1] or adaptive synthetic sampling approach. ADASYN works by generating synthetic data for the minority class in a way that prioritizes harder-to-learn examples. It assigns a weighted distribution to minority samples based on their difficulty, producing more synthetic instances for those that are harder to classify while generating fewer for easier cases. This helps improve the performance of the neural network by ensuring it does not become biased toward the majority class. Figure 4 highlights the effectiveness of the ADASYN approach, particularly for classes 4 and 5, where high recall is crucial due to their severity as malignant cases. With ADASYN, recall for class 4 improves from 10% to 36% and for class 5 from 0% to 40%, preventing the model from misclassifying nearly all samples as class 3, the majority class. To sum up, the best results (F1-score) we could obtain for this classification were as following; 30%, 30%, 43%, 32%, 27% for classes 1,2,3,4,5 respectively and overall macro F1-score of 33%. A lot of work was done focusing on feature selection. The idea was to reduce the noise (irrelevant and redundant features) and improve efficiency in training. We tried three different methods. The first was to look at which features were highly correlated and within each correlated pair, keeping the feature with higher feature importance. After removing those with an 85% threshold, we were left with 1145 features. The second was to perform SHAP analysis. By doing so, features were removed iteratively and manually based on their SHAP values and with checking the corresponding classification results. Finally, the third method considered was to perform PCA and maintain only the components which accounted for a certain proportion of the variance. However, these feature reduction techniques may not always (and didn't) improve performance due to loss of relevant information, ignored non-linear interactions, or reduced interpretability.



(a) Before ADASYN (b) After ADASYN

Figure 4: Comparison of the classification results

5. Explainability

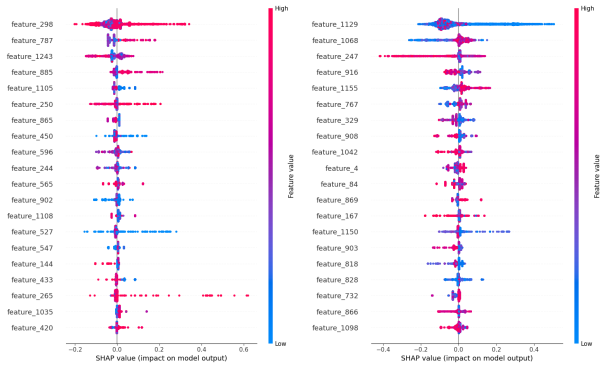
Neural networks are extremely powerful tools and they are well suited for a wide variety of tasks, including the task at hand of classifying lung tumor severity. However, they remain black-box models and as such they provide us with an output without any explanation of how they got to this output. As such, in class we saw the importance of applying explainability measures to these black-box models. They allow us to glean some insights on how decisions are made by the model, for example by showing us which features are more important than others, or which section of the image it uses to make a decision. In the medical context to which our problem belongs, this explainability is all the more crucial. It means that clinicians who use our model can assess how the model is making decisions and this allows them to trust it and use it effectively. Given the delicate and important nature of explainability tools, we tried a few different ones. Given the length constraint of the report, we do not show examples of the explainability measures that we talk about to all the models in the report, but rather provide a few examples. Specifically, we use Shapley values on the full slice model and grad cam on the nodule model.

5.1. SHAP

SHAP (Shapley Additive Explanations) quantifies the impact of each feature on the target variable, based on Shapley Values from Game Theory. In this context, features act as players, and the dataset is the team—SHAP measures each feature's contribution to the prediction. It calculates SHAP values using combinatorial calculus by retraining the model on all possible feature combinations. The average absolute SHAP value serves as a metric for feature importance. Here, SHAP was applied to the full slice 5-class problem to analyze the contribution of different extracted features to malignancy classification, and the results of the analysis on our classifier are reported below.

In our case, positive SHAP values indicate that a feature pushes the prediction toward a higher malignancy class, whereas negative SHAP values push it toward a lower malignancy class. The first plot,

averaging all five malignancy classes, highlights **feature_1129, feature_751, and feature_298** as the most influential, with high SHAP values indicating a strong impact on model predictions. The second plot, focusing on class 4 (high malignancy), shows **feature_298 and feature_787** playing key roles, with SHAP values distributed across positive and negative ranges, reflecting complex interactions. The third plot, representing class 5 (most malignant), highlights **feature_1129 and feature_1068** as dominant, with a wider SHAP value distribution, indicating substantial variability in feature importance. These results suggest that specific features consistently influence malignancy predictions, though their impact varies across severity levels, providing insights into how the model differentiates between different malignancy classes.



(a) Class 4 (b) Class 5

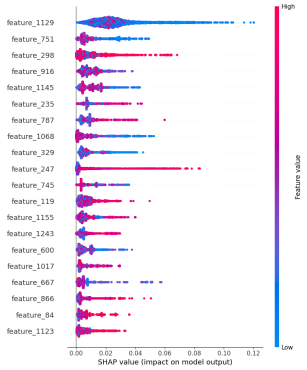


Figure 6: All classes

5.2. Grad-CAM

Grad-CAM is a very useful tool for explainability as it highlights which parts of the image have been used by a convolutional neural network to make its decision. It is a model-specific method that provides the user with local explanations. It calculates gradients of the target class score with respect to the feature maps in the last convolutional layer, averaging these gradients to determine the importance of each feature map. Red / yellow areas are the most important regions for the model's prediction, while blue / purple Area are less relevant to the decision. Below we see Grad-CAM applied to the images of the binary classifier for the nodules. There are two images because the binary classifier for the nodules uses two pretrained models: InceptionResNet and DenseNet and thus it is applied to the last convolutional layer extracted from both models.

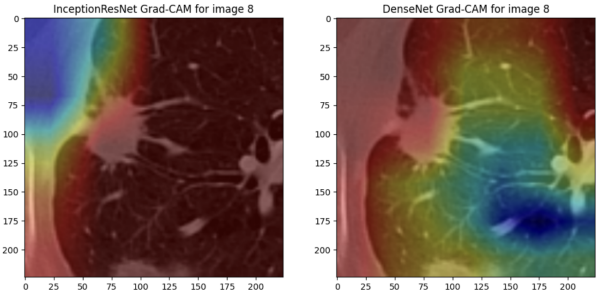


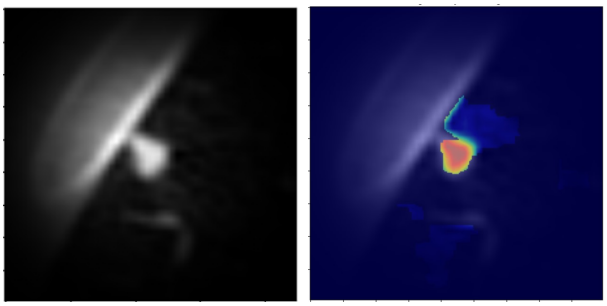
Figure 7: Grad-CAM no smoothing

What applying Grad-CAM revealed is that the model is not always focusing on the nodule, but sometimes it focuses on the sorrounding areas, like the bone, which do not help identify if a tumor is malignant or not.
This gave us the idea to perform a smoothing at the edges of the images, by applying a Gaussian filter to the input images (see section 3.1). The idea is to direct the model to focus on the center of the image, where the nodule is usually contained.

5.3. LIME

LIME (Local Interpretable Model-agnostic Explanations) is a powerful tool for explainability that helps users understand how machine learning models make their predictions. It is a model-agnostic method that provides local explanations by approximating the complex model with an interpretable surrogate model. LIME perturbs the input data and observes how the model's predictions change, fitting a simple model (such as a weighted linear regression) to capture the decision boundary in the vicinity of the instance being explained. This allows users to identify which features were most influential in a particular prediction. A heatmap is created by weighting the feature maps and applying ReLU, showing relevant areas for the prediction. The color of an area of the image then provides an indication of how important that area was for the models decision. Red / yellow areas are the most important regions for the model's prediction, while blue / purple Area are less relevant to the decision.

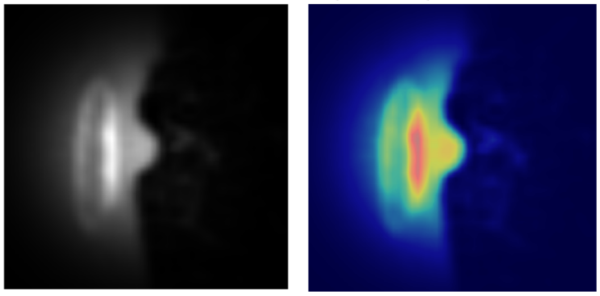
Let's try applying LIME to a nodule image from the test set that was correctly classified as a benign nodule by the 2-classes model:



(a) Input + filtering (b) Explanation

Figure 8: Comparison of the original image and LIME explanation

As we can see, the model correctly focuses only on the nodule, and therefore classifies it correctly as benign. Let's see another example, this time from the 5-classes nodule model. From the confusion matrix, we can observe that one sample from the test set belongs to class 4 but is misclassified as class 1. We care a lot about this example, cause misclassifying a stage 4 cancerous nodule as stage 1 is a critical and potentially life-threatening error. Now, let's examine the LIME output for this sample:



(a) Input + filtering (b) Explanation

Figure 9: Comparison of the original image and LIME explanation

As we can see, the model is not focusing on the nodule, but rather on the center-left part of the image, close to the bones, therefore it misclassifies the sample.

6. Conclusions

These results highlight the current limitations of machine learning models in independently performing such a critical medical task. The risk of misclassification, especially in cases where a life-threatening condition is mistaken for a less severe one, makes it ethically unjustifiable to rely solely on automated systems. However, explainability offers a potential solution by allowing experts to understand how and why a model reaches its decisions. By making the decision-making process more transparent, techniques such as LIME and Grad-CAM enable medical professionals to verify whether the model is focusing on the correct features, identify potential biases, and intervene when necessary. Rather than replacing human expertise, explainable AI can serve as a valuable tool to assist specialists, ensuring that automated systems support, rather than compromise, patient safety.

7. Bibliography and citations

References

[1] Haibo He and Yang Bai and Garcia, Edwardo A. and Shutao Li *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), ADASYN: Adaptive synthetic sampling approach for imbalanced learning* `https://ieeexplore.ieee.org/document/4633969`

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, *"Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.* `https://arxiv.org/pdf/1610.02391`

[3] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* `https://arxiv.org/pdf/1602.04938`