Prediction of Malignancy in Lung Cancer using several strategies for the fusion of Multi-Channel Pyradiomics Images

Jan Rodríguez Dueñas

Resum– Aquest treball mostra el procés de generació i el posterior estudi de l'espai de representació que obtenim extraient característiques de textura GLCM de tomografies assistides per ordinador (CT) dels nòduls pulmonars (PN). El flux del treball se centra en l'extracció de característiques mitjançant Pyradiomics i la xarxa neuronal convolucional (CNN) VGG16. L'estudi té com a objectiu valorar si les dades aconseguides impacten de manera positiva en el diagnòstic de càncer pulmonar (LC). Per dissenyar un mètode d'entrenament de models d'aprenentatge automàtic (ML) que permet generalització, entrenem models SMV amb diferents divisions de dades, valorant el rendiment de la diagnosi mitjançant mètriques definides a nivell de tall i de nòdul. Per aquesta tasca, s'han utilitzat dades de 92 pacients de l'Hospital Universitari Germans Trias i Pujol.

Paraules clau– Càncer pulmonar, diagnostic precoç, cribratge, radiòmica, espai de representació, SVM, optimització de models.

Abstract– This study presents the process of generation and subsequent study of the representation space obtained by extracting GLCM texture features from computed tomography (CT) of pulmonary nodules (PN). The workflow of the study focuses on feature extraction using Pyradiomics and the VGG16 convolutional neural network (CNN). The objective of the study is to assess whether the obtained data have a positive impact on the diagnosis of lung cancer (LC). To design a method for training machine learning (ML) models that enables generalization, we train SVM models with different data splits, evaluating the diagnostic performance using metrics defined at both the slice and nodule levels. For this task, data from 92 patients from the Hospital Universitari Germans Trias i Pujol have been used.

Keywords– Lung cancer, early diagnosis, screening, radiomics, representation space, SVM, model optimization.

1 Context

Lung cancer (LC) is the leading cause of cancer-related mortality worldwide. Its impact is profound, with millions of lives affected and a significant burden on healthcare systems. LC arises from the uncontrolled growth of abnormal cells in the lung tissues, leading to the formation of tumors.

Early detection plays a crucial role in improving treatment outcomes and patient prognosis. Studies such as the National Lung Screening Trial (NLST) [1] and NELSON [2] have demonstrated that annual screening with low-dose computed tomography (LDCT) can effectively reduce mortality rates associated with LC [3].

However, LC screening presents challenges that include the need for further imaging and follow-up procedures in cases of positive findings. These additional investigations can lead to patient anxiety due to the potential invasiveness of the procedures and impose significant costs on healthcare services.

Fortunately, advances in the field of radiomics [4] have revolutionized lung cancer screening and management. Radiomics involves the extraction of a large number of quantitative features from medical images, such as computed tomography (CT) scans, magnetic resonance imaging (MRI), or positron emission tomography (PET). These fea-

• E-mail de contacte: jan.rodriguez@autonoma.cat
• Treball tutoritzat per: Débora Gil Resina (Ciències de la Computació)
• Curs 2022/23

tures capture the heterogeneity and characteristics of lung tumors at a microscopic level, enabling more precise diagnosis and treatment planning.

In a pilot study [5], image features extracted from NLST (National Lung Screening Trial) data using radiomic techniques exhibited superior predictive value compared to volumetric measurements alone. This highlights the potential of radiomics in enhancing lung cancer screening accuracy. Additionally, Peikert [6] developed a radiomic classifier incorporating location variables, size, shape descriptors, and texture analysis. This innovative approach further demonstrates the power of radiomics in providing valuable insights for accurate lung cancer detection and characterization.

The integration of radiomics with genomics, metabolomics and clinical data holds promise in deciphering the complex biology of lung cancer [7]. This multidimensional approach allows for a comprehensive understanding of tumor behavior, treatment response, and patient prognosis. By combining radiomic features with genomic profiles, researchers can identify potential therapeutic targets and explore personalized treatment options.

The RADIOLUNG project is a multicentric initiative coordinated by the Interactive Augmented Modeling (IAM) group at the Computer Vision Center (CVC) in collaboration with Hospital Universitari Germans Trias i Pujol (HUGTiP).

The primary objective of the RADIOLUNG project is to develop a comprehensive multi-radiomic signature based on chest CT and PET-scan images for distinguishing benign and malignant pulmonary nodules (PNs). Additionally, the project aims to evaluate the predictive capabilities of this signature and assess whether it can reduce the false-positive rate by more than 50

The specific objectives of the RADIOLUNG project are as follows:

1. Correlate the pathological and molecular profiles with each radiomic signature and investigate the presence of clinically relevant mutations in tumors.

2. Determine the degree of aggressiveness and mutational status of PNs, and explore the possibility of identifying an epigenetic profile associated with malignancy using radiomic signatures.

3. Integrate imaging, genomics, and clinical data to develop a predictive model for nodules detected through low-dose computed tomography.

4. Design a training methodology that enables generalization of the predictive model across multiple centers.

5. Optimize the architecture of the predictive model to classify lung nodules into three diagnostic levels.

By achieving these objectives, the project aims to contribute to the advancement of LC screening and pave the way for more accurate and efficient early detection strategies. The collaboration between the IAM group, the CVC and HUGTiP provides a strong foundation for this research, ensuring access to expertise and resources necessary for successful completion, ultimately benefiting the field of lung cancer screening and improving patient outcomes.

This TFG is closely aligned with the RADIOLUNG project. While the RADIOLUNG project has its own comprehensive objectives, this TFG complements the project by focusing on the reduction of the false positive rate by the analysis of CT scans.

2 State of the Art

The effectiveness of lung cancer screening in reducing mortality is hindered by a high rate of false positive results, scarcity of data, and rare occurrence of benign cases. Deep learning methods, despite being state-of-the-art, can be problematic due to bias, overfitting, and lack of reproducibility. In contrast, machine learning approaches [8, 9] utilizing established techniques like Gabor, Local Binary Patterns (LBP), and SIFT descriptor, combined with classifiers such as Support Vector Machine (SVM) and Random Forest, have shown improved diagnostic power with high sensitivity and specificity, achieving an AUC of 0.97 and sensitivity of 96% with 95% specificity for [8].

GLCM (Gray-Level Co-occurrence Matrix) texture features, have demonstrated effectiveness in cancer diagnosis across various medical imaging modalities [10, 11, 12, 13, 14]. In a recent study [15], researchers proposed a hybrid approach that combined GLCM textural features with a neural network for nodule characterization in CT scans. To ensure reproducibility with limited training data, an embedding technique based on the statistical significance of radiomic features was used. This embedded representation served as the input for a neural network, with its architecture and hyperparameters optimized using custom-defined metrics. The best performing model achieved a sensitivity of 100% and specificity of 83% (with an AUC of 0.94) for malignancy detection when evaluated on an independent patient set. This innovative approach shows promise in improving the accuracy and reliability of lung cancer screening by integrating radiomic features and deep learning techniques, offering potential solutions to the challenges posed by false positives in current screening methods.

3 TFG Objectives

The analysis of the current State of the Art methods indicates that the visual embedding representation is a crucial step for the correct diagnosis of the PN. This project analyzes the benefits of combining deep features with radiomic texture features.

The specific objectives of this TFG are:

1. **Generate a representation space based on deep features (specifically Vgg16) extracted from GLCM texture images.** To study the benefits of combining deep features with radiomic texture features, a representation space based GLCM texture images is computed.

2. **Generate a representation space based on instensity deep features using Vgg16.** In order to study the benefits of using texture images, a representation space based on the intensity images has also been computed to do a comparison between both methods.

3. **Optimize an SVM for the classification of malignancy.** To do the classification of the malignant nodules and study the impact of every representation space, diagnosis performance metrics will be computed with the predictions made by an SVM model. Those are the same metrics that we will check to decide which is the most optimal model.

4. **Compare the detection of malignancy of the 2 representations spaces at 3 different levels of generalization.** Additionally, this project explores three different levels of generalization: nodule k-fold, leave-1-nodule-out, and slice k-fold. These levels provide further insights and enable us to evaluate the performance and robustness of our methodology.

4 METHODOLOGY

In this Section, we explain the main steps of the proposed strategy for malignancy detection based on deep textural features extracted using texture images and VGG16 (Section 4.1), as well as, the validation protocol for the assessment of its level of generalization (Section 4.2). The rest of this section is dedicated to give more detailed explanations of each stage.

4.1 Strategy for Diagnosis of Malignancy

Our workflow consists of multiple steps, which are illustrated in Figure 1. First, we extract the nodule region of interest (ROI) from CT scans using a predefined ROI. Subsequently, in the generation of the representation space defining a visual embedding of the nodule, we have the option to either pass the ROI without any modifications or extract GLCM features from it. These features are then fed into a pre-trained VGG16 network to obtain the final feature embeddings. We have explored three different strategies for feature fusion to combine these embeddings and train a model for predicting the axial 2D images of the nodule ROI. The final nodule diagnosis is determined using a max-voting criteria.

4.1.1 Nodule Extraction

In our workflow, we begin receiving anonymized CT-chest scans in DICOM format, which are then converted to the NIFTI format. The NIFTI format is specifically designed for neuroimaging data and proves to be highly suitable for preprocessing tasks. Afterward, a radiologist defines 3D bounding boxes in the CT scans that encompass the nodules. These bounding boxes serve as references for extracting the Regions of Interest (ROIs) from the CT scan, which are subsequently utilized throughout the remainder of our workflow. We refer to these extracted ROIs as nodule ROIs, as they represent the outcome of this preprocessing step. The manual annotation of these 3D bounding boxes is the only instance of human intervention required.

Regarding the size of the ROI, there are two important aspects to consider. First, the nodule ROI always includes the intranodule region (the nodule itself), but the extent of the perinodular region (the area around the nodule) varies depending on the shape of the nodule. Studies such as [16, 17] have highlighted the importance of including the perinodular region in accurately classifying benign and malignant nodules. Therefore, we increase the size of the ROIs to encompass this aspect. Second, in a subsequent step of our workflow, we utilize a VGG16 network that requires a minimum input image size of 42x42 pixels (width and height). If any of the extended nodule ROIs are below this minimum size, we further expand them to meet the network's requirement.It is worth noting that nodule ROIs may have different widths and heights due to variations in ROI sizes.

4.1.2 Nodule Embedding

We employ two methods to generate the representation space using the nodule ROIs that were extracted in section 4.1.1 as is depicted in Figure 1. The goal of this feature embedding step is to derive meaningful and discriminative representations of the nodules, which can be further analyzed and used for classification tasks.

The GLCM textural features are calculated using the nodule ROI. Additionally, for each nodule, we generate a fictitious nodule mask where all voxel values are set to one. This mask indicates that all voxels within the nodule ROI should be considered when computing the GLCM features. By employing this nodule mask, we can generate 21 GLCM features (i.e., 21 volumes) for each nodule ROI, corresponding to the textural features computed in [15].

Let us provide a more in-depth understanding of this approach. GLCM features are statistical descriptors computed from a gray-level co-occurrence matrix. This matrix captures the frequency of occurrence of pixel pairs with specific gray-level values and spatial relationships within a defined neighborhood.

To generate the GLCM features, we begin by discretizing the intensity gray values using the histogram of the original volume intensity. This process involves dividing the range of gray values into discrete bins. The width of these histogram bins determines the level of granularity at which the GLCM features describe the textural patterns. Smaller bin widths provide a finer level of detail, while larger bin widths result in more generalized information.

Once the gray values are discretized, the GLCM is constructed by examining the spatial relationships between pixels within the neighborhood. Specifically, for each pixel, the occurrence of gray-level pairs and their spatial relationships with neighboring pixels are recorded in the co-occurrence matrix.

Based on the GLCM, a variety of statistical measures can be calculated to extract textural information. These measures include contrast, correlation, energy, homogeneity, and many others. Each measure provides insights into different aspects of texture, such as the variations in pixel intensities, the degree of similarity between neighboring pixels, and the overall uniformity of the texture.

In summary, the GLCM features offer a way to quantify and characterize textural patterns within the nodule ROI. By examining the statistical relationships between pixel pairs, these features provide valuable information for cancer diagnosis and have been widely employed in various medical imaging modalities.

To extract deep features, axial slices of both the original intensity volume and the 21 GLCM texture volumes are the

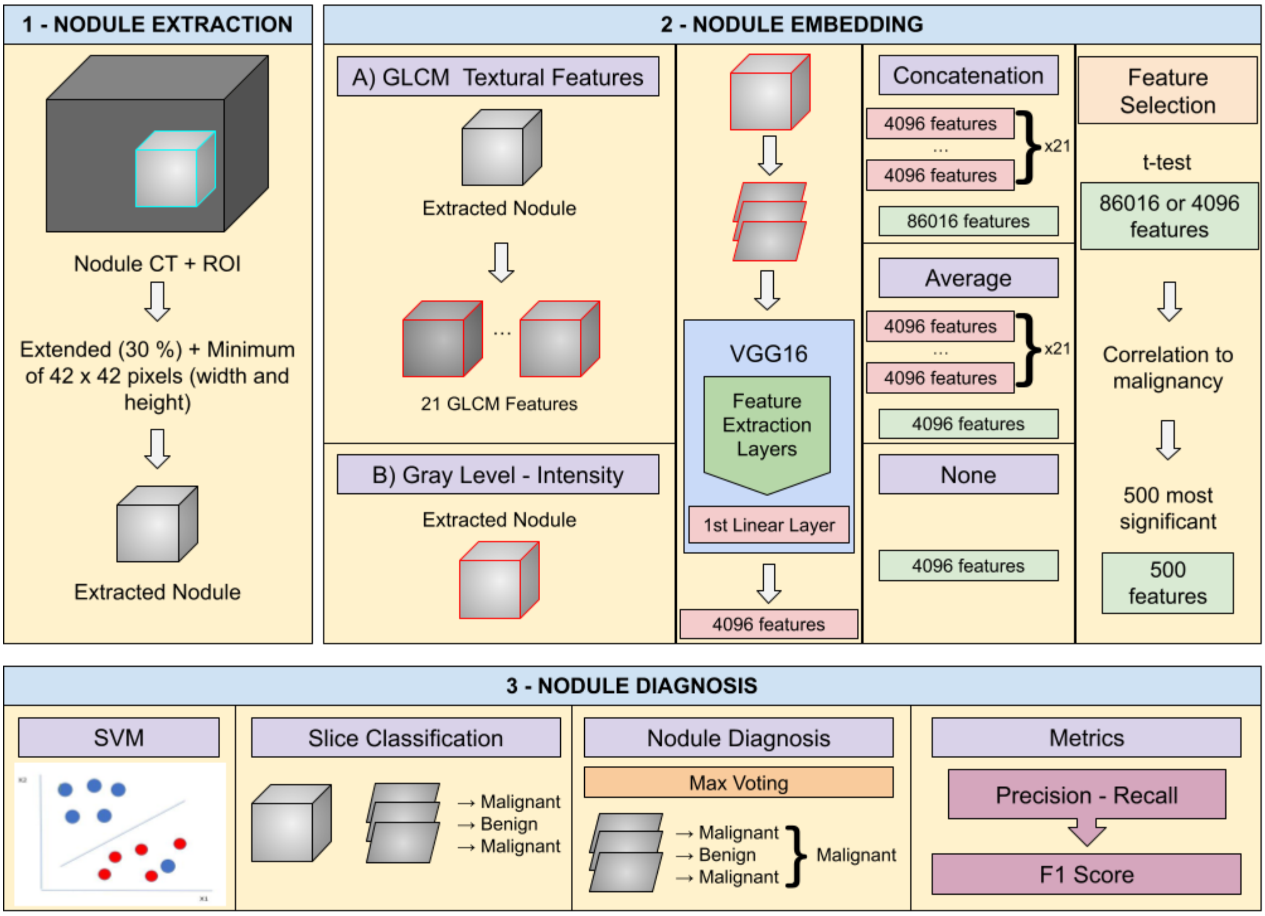

Fig. 1: Workflow of the strategy for diagnosis of malignancy.

input for a pre-trained VGG16 model that has been trained on ImageNet [18].

The VGG16 architecture is composed of 13 convolutional layers, 5 max-pooling layers (2 × 2), and 2 fully-connected layers. The linear output layer utilizes the softmax activation function. ReLU activation function is applied to all the convolutional layers, while dropout regularization is employed in the fully connected layers. The deep representation for both intensity and texture images is defined by the features extracted from the FC6 layer.

For each image, the deep feature vector from the FC6 layer has a dimensional size of 4096. In the case of intensity images, this results in 4096 features. However, for the GLCM features approach, which includes 21 GLCM volumes per nodule, the resulting features have a dimension of (21, 4096). These 21 channels need to be combined to create an input vector for a classifier. There are three options considered: concatenation, average, and none.

In the concatenation strategy, the features are flattened, resulting in 86016 features. This means that the 21 channels are concatenated to form a single long feature vector.

In the average strategy, the features are used to compute an average. This results in a single feature vector with the same dimension as each individual feature vector, i.e., (1, 4096).

Lastly, for the raw gray levels features, they result in features with a dimension of (1, 4096), meaning that there is only one channel in the feature vector.

Regardless of the chosen strategy for features fusion, we proceed to apply a t-test to rank the features based on their significance in correlating with nodule malignancy. This step enables us to perform feature selection and identify the most relevant features. For the VGG features, they are ranked based on the p-value obtained from a t-test that measures the difference in averages between malignant and benign slices. The top 500 features with the lowest p-values are then selected as input for the SVM classifier.

4.1.3 Nodule Diagnosis

We employ the fused features (none, concatenation, and average) obtained from Section 4.1.2 to train an SVM classifier for making slice-by-slice predictions.

To optimize the SVM parameters, we perform a grid search method where multiple combinations of the parameters C, kernel, and gamma are tested. This process helps identify the parameter settings that yield the most favorable outcomes.

After training the SVM, the diagnosis of the nodule is determined by aggregating the slice predictions using a max-voting approach. This approach considers the most frequent slice classification to determine the final diagnosis of the nodule. For example, if more than half of the 2D slices are classified as malignant, the overall diagnosis is considered malignant. Otherwise, it is classified as benign. In the case of a tie, we assign a malignant diagnosis.

4.2 Levels of Generalization

We split the data into three levels of generalization to study the impact of the new representation space. It is important to note that the experimental unit differs between the approaches. For nodule k-fold and leave-1-nodule-out, the experimental unit is the nodule itself. This means that the slices of a nodule can only be in either the training or the test set, but not both. On the other hand, the experimental unit for the slice k-fold approach is the individual slice, allowing slices from the same nodule to be present in both sets. These approaches provide valuable insights into the model's performance and generalization capabilities, ensuring a robust evaluation with a high degree of generalization and reproducibility.

1. Nodule k-folds: in this approach, we employ k-fold cross-validation to assess the performance of our model when predicting on unseen data. The unit of data splitting is the nodule, where the dataset is divided into subsets of nodules. One subset is designated as the test data, while the remaining subsets are used for training. This process is repeated k times, with each fold using different nodules for testing. After training a model for each fold, the diagnosis score is computed by averaging the performance across the k-folds. This approach provides two levels of measures: individual fold performance and an overall measure of the model's performance using all folds. Thus, we can capture statistical ranges at each fold and across all folds.

2. Leave-1-Nodule-Out: This approach represents a particular implementation of k-fold cross-validation, as explained earlier, with k set to the maximum number of nodules in the dataset. The nodule serves as the experimental unit for data splitting. Accordingly, the subsets of nodules consist of one nodule assigned to the test data, while the remaining nodules form the training data. Since the test set contains only one nodule, this approach yields a single level of measurement, which is an overall evaluation of the model's performance using all folds. Consequently, statistical ranges can be captured across all folds.

3. Slice k-folds: In this approach, we utilize k-fold cross-validation, with the slice as the experimental unit. This means that slices from the same nodule can be present in both the training and test data. The process is repeated k times, with each fold incorporating different slices for testing. The diagnosis score is calculated as the average across the k-folds. Similar to the nodule k-folds method, this approach captures statistical ranges at each fold and across all folds.

5 Experimental Set-up

This study utilizes our database, which comprises patients recruited from the Germans Trias i Pujol University Hospital (HUGTiP) in Barcelona, Spain. The database includes images and clinical/demographic data collected between December 2019 and November 2022. A total of 92 patients were included in this study. All patients underwent low-dose CT-chest scans and had pulmonary nodules (PN) that required surgical intervention. The selection of patients was based on specific inclusion and exclusion criteria. Inclusion criteria required patients to have a single PN with a diameter ranging from 8 to 30 mm and a confirmed diagnosis of either non-small cell lung carcinoma or a non-malignant tumor. Exclusion criteria included a previous diagnosis of lung cancer, the presence of incurable extrapulmonary cancer (excluding non-melanoma skin cancer), pregnancy, recent chemotherapy or cytotoxic drug use within the last 6 months, and refusal to provide informed consent. Notably, each PN underwent a biopsy procedure to accurately determine its pathological nature.

Table 1 provides detailed information about our database, including demographic data, the range of slices for each nodule type and sex. Our database comprises a total of 92 lung nodules, which were used for conducting our experiments.

The minimum size of a nodule region of interest (ROI) is 42x42 pixels (width and height), as imposed by the pretrained VGG16 network described in section 4.1.1. For the computation of the 21 GLCM features [19], we utilized PyRadiomics [20] (version 3.01). Since the GLCM features are calculated using the nodule mask, which identifies the specific voxels to be included in the computation, a fictitious mask was created for each nodule, containing all ones. This ensures that the generated GLCM features match the size of the nodule ROIs, and also preserve the perinodular region. Specifically, a (3 × 3 × 3) kernel was used to determine the voxels involved in the calculation of GLCM features, and the image was discretized into 128 bins.

The computation of GCLM features was a significant bottleneck in the workflow, as it took more than 40 hours to complete. To address this issue, we implemented a solution to parallelize the calculation using Parfor [21] and made use of a cluster with 60 CPUs. This optimization dramatically reduced the execution time to less than 2 hours. To quantify the improvement achieved by the parallel execution compared to sequential execution, we measure the speed-up performance as follow:

$$Speedup(n) = \frac{T(1)}{T(n)}$$

being n the number of CPUs used. Similarly, efficiency measures the utilization of computational resources and is computed as follow:

$$E(n) = \frac{S(n)}{n} = \frac{T(1)}{nT(n)}$$

After extracting the features, we apply a t-test for each feature fusion method, such as concatenation, average, or no fusion. The t-test allows us to obtain a rank of the most significant features. From this ranking, we select the top 500 most important features to train an SVM classifier. This selection process takes into account different levels of generalization, as discussed in Section 4.2, to study the generalization and reproducibility of our method as follow:

1. Nodule 5-folds: The nodules were split into folds using the StratifiedGroupKFold function from the scikit-learn Python package. The split was done in a way that maintained the proportion of classes within each fold. With a value of k set to 5, approximately

TABLE 1: DEMOGRAPHIC INFORMATION OF OUR DATABASE

	Description	Male	Female	Total
Demographic Population	Patients	63	29	92
	Age avg ±std	74 ±7	69 ±11.4	73 ±9
	Benign PNs	8	5	13
	Malign PNs	55	24	79
Nodule characterization	Benign Slices min/max/avg	6/111/41.4	28/50/36.2	6/111/39.3
	Malign Slices min/max/avg	8/152/45.9	12/105/47.4	8/152/46.3

18 or 19 nodules were included in each fold. The diagnosis score is computed as the average performance in each fold individually and across the 5 folds, providing a reliable estimation represented as a confidence interval.

2. Leave-1-Nodule-Out: In this approach, k is set to 92, which is the same number of nodules in our database. Each fold consists of leaving one nodule out as the test set, while the remaining nodules are used for training. The diagnosis score is then computed based on the predictions made across the 92 folds (test folds), providing a single global measure.

3. Slice 5-folds: For this method, k is set to 5. Each fold consists of 277 slices used as the test data, while the remaining slices are utilized for model training. It is important to note that different slices from the same nodule can be present in both the training and test sets. This process is repeated five times, with different slices used for testing in each fold. The diagnosis score is then calculated as the average in each fold individually fold and across the five folds and represented as a confidence interval. For this experiment, 77 ($\approx$ 73%) nodules of the dataset were randomly selected for the training of the models. In this way, the independent set (Holdout) of test patients is conformed by a total amount of 25 nodules with 7 benign and 18 malign. This holdout acts as an independent set to test our models and evaluate their performance. It provides an unbiased estimate of how well the model would generalize to unseen data

To compare the results obtained for each evaluation method, with malignant nodules considered as positive cases, we computed the following metrics based on true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) at both the slice and nodule levels:

$$F1 - score = 100 * \frac{2 * Prec * Rec}{Prec + Rec}$$

or Rec, Prec denoting, respectively, the precision and recall at diagnosis level:

$$Rec = 100 * \frac{TP}{TP + FN}$$

$$Prec = 100 * \frac{TP}{TP + FP}$$

Precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of relevant instances that were successfully retrieved. These metrics help us measure the performance of our classifier in terms of false positives and false negatives. The F1-score combines precision and recall into a single value, representing the trade-off between the two metrics. A higher F1-score indicates better overall performance.

6 RESULTS

The parallelization of the feature extraction process using 60 CPUs makes us calculate the GLCMs of the nodules in 1h and 36 minutes (1.6 hours). With a time of 40 hours for 1 CPU, that means that the speed-up achievied is

$$S(60) = 40/1.6 = 25$$

In consequence, the efficiency of our sistem, that is defined by the speed-up and the number of CPUs used, is

$$E(60) = 25/60 \approx 0.42$$

In regards to the experiments related to feature space and diagnosis, they were conducted with the aim of comparing the performance of classifiers depending on the data they were trained on.

The obtained results from the optimal configurations are presented in Table 2, illustrating the SVM classifier's performance in terms of precision, recall, and F1-score at the nodule level.

Regarding the Diagnosis score, it is observed that the Intensity domain has the lowest score among all domains. When using slice folds for splitting, both GLCM-Concatenation and GLCM-Average domains exhibit high recall for both benign and malignant nodules. The recall range for GLCM-Concatenation is (1, 1) for malignant cases and (0.84, 1) for benign cases. However, when splitting at the nodule level, the GLCM-Average domain experiences a significant drop in benign recall, almost reaching 0. On the other hand, for the GLCM-Concatenation domain, while the malignancy recall score falls within the range of (0.88, 1), the recall range for benign cases is (0.37, 1). It is worth noting that the high standard deviation (around 30%) indicates considerable variability across folds for the GLCM-Concatenation domain. This variability can be attributed to the limited number of benign samples, with only 1, 2, or 3 samples at most. As a result, a false positive result can lead to a recall variation of 33%, 50%, or even 100%.

TABLE 2: Diagnosis score at Nodule Level in experiments with SVM

Data Domain	Split	Diagnosis	Precision	Recall	F1-score
Intensity	Nodule 5-folds	Malign	0.85 ± 0.04	1.00 ± 0.00	0.92 ± 0.02
		Benign	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	L10	Malign	0.86	1.00	0.92
		Benign	1.00	0.07	0.13
	Slice 5-folds	Malign	0.95 ± 0.01	1 ± 0.00	0.98 ± 0.01
		Benign	1 ± 0.00	0.49 ± 0.09	0.65 ± 0.08
	Holdout	Malign	0.69	1.00	0.82
		Benign	0.00	0.00	0.00
GLCM-Concatenation	Nodule 5-folds	Malign	0.94 ± 0.05	0.94 ± 0.06	0.94 ± 0.04
		Benign	0.70 ± 0.27	0.67 ± 0.30	0.63 ± 0.22
	L10	Malign	0.90	0.95	0.93
		Benign	0.56	0.39	0.46
	Slice 5-folds	Malign	1 ± 0.00	0.99 ± 0.01	0.99 ± 0.00
		Benign	0.91 ± 0.07	1 ± 0.00	0.95 ± 0.04
	Holdout	Malign	0.75	0.83	0.79
		Benign	0.40	0.29	0.33
GLCM-Average	Nodule 5-folds	Malign	0.87 ± 0.04	1.00 ± 0.0	0.93 ± 0.02
		Benign	0.20 ± 0.40	0.10 ± 0.20	0.13 ± 0.27
	L10	Malign	0.86	1.00	0.92
		Benign	0.00	0.00	0.00
	Slice 5-folds	Malign	0.99 ± 0.01	1 ± 0.0	1 ± 0.00
		Benign	1 ± 0.0	0.93 ± 0.09	0.96 ± 0.05
	Holdout	Malign	0.72	1.00	0.84
		Benign	1.00	0.14	0.25

7 Conclusions

While the evaluation of parallelization and its impact on system efficiency was not initially a part of our original objectives, the results obtained highlight its significant effect on speeding up the feature extraction process. By utilizing 60 CPUs, we were able to calculate the GLCMs of the nodules in just 1 hour and 36 minutes, achieving a remarkable speed-up of 25 compared to a single CPU. The resulting efficiency of ≈ 0.42 demonstrates the successful utilization of parallel computing.

Although the enhancement of system performance through parallelization was not a primary objective, it is worth noting that the presence of a larger number of nodules would likely further improve efficiency. The current bottleneck lies in processing larger-sized nodules, which are more time-consuming. By increasing the quantity of nodules and distributing the workload more evenly, the system's performance could be optimized even further.

As shown in Table 2, the domain that exhibits the lowest performance is Intensity. This can be attributed to the fact that VGG16, which was trained on ImageNet for object classification in natural scenes, may not effectively capture the texture details characteristic of cancer tumor lesions. On the other hand, GLCM demonstrates higher discrimination power as it can represent the texture details of the nodules.

Regardless of the representation space, the data split at the slice level yields the least reproducible results. This is because the intervals do not contain the metrics of the hold-out independent test set. Additionally, they present overly optimistic precision, recall, and F1-score values for malignant nodules within the range of (0.98, 1), and (1, 1) for benign nodules, when using the GLCM-Concatenation domain. These numbers are comparable to, or even better than, those achieved by state-of-the-art methods. However, the metrics for the hold-out set drop to 0.83 for malignant nodule recall and 0.29 for benign nodules. The interval predictions obtained by splitting the data at the nodule level are less optimistic but more realistic as they include the hold-out metric results.

These observations highlight the challenges and limitations in achieving consistent and reliable results in lung cancer screening. The findings underscore the need for further research and development to address issues related to dataset size, imbalance and reproducibility, ultimately improving the accuracy and reliability of screening methods.

Acknowledgements

I would like to begin by expressing my sincere gratitude to all the individuals who have contributed to the completion of my TFG. Your support and collaboration have been fundamental throughout this research and writing process.

First and foremost, I would like to extend my thanks to everyone who has been by my side during this exciting academic journey. Your presence, encouragement, and constant support have been a driving force for me.

I want to express my deep appreciation to the team at IAM (Institute of Mathematical Analysis), especially my supervisor, Débora. I am immensely grateful for her dedication, guidance, and expert knowledge. Her mentorship has been instrumental in the development and success of my TFG.

Special thanks go to Guillermo and Carles from IAM for their technical support whenever I needed it. Their expertise and assistance have been vital in overcoming technical challenges and advancing my research.

I am grateful to the Computer Vision Center for allowing me to work with one of their teams in their facilities. I appreciate their generosity and trust in providing me with the resources and conducive environment for the completion of my TFG.

In conclusion, I would like to once again recognize all the individuals mentioned above, as well as anyone else who has contributed in any way to my TFG. Their collaboration has been essential to this academic achievement.

References

[1] N. L. S. T. R. Team, "The national lung screening trial: overview and study design," *Radiology*, vol. 258, no. 1, pp. 243–253, 2011.

[2] H. J. de Koning, C. M. van der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg *et al.*, "Reduced lung-cancer mortality with volume ct screening in a randomized trial," *New England journal of medicine*, vol. 382, no. 6, pp. 503–513, 2020.

[3] C. for Disease Control and Prevention, "Who should be screened for lung cancer?" 2022, accessed June 29, 2023. [Online]. Available: https://www.cdc.gov/cancer/lung/basic_info/screening.htm

[4] J. D. Shur, S. J. Doran, S. Kumar, D. Ap Dafydd, K. Downey, J. P. O'Connor, N. Papanikolaou, C. Messiou, D.-M. Koh, and M. R. Orton, "Radiomics in oncology: a practical guide," *Radiographics*, vol. 41, no. 6, pp. 1717–1732, 2021.

[5] Y. Liu, J. Kim, Y. Balagurunathan, S. Hawkins, O. Stringfield, M. B. Schabath, Q. Li, F. Qu, S. Liu, A. L. Garcia *et al.*, "Prediction of pathological nodal involvement by ct-based radiomic features of the primary tumor in patients with clinically node-negative peripheral lung adenocarcinomas," *Medical physics*, vol. 45, no. 6, pp. 2518–2526, 2018.

[6] T. Peikert, F. Duan, S. Rajagopalan, R. A. Karwoski, R. Clay, R. A. Robb, Z. Qin, J. Sicks, B. J. Bartholmai, and F. Maldonado, "Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the national lung screening trial," *PLoS One*, vol. 13, no. 5, p. e0196910, 2018.

[7] G. Lee, H. Y. Lee, H. Park, M. L. Schiebler, E. J. van Beek, Y. Ohno, J. B. Seo, and A. Leung, "Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art," *European journal of radiology*, vol. 86, pp. 297–307, 2017.

[8] F. Zhang, Y. Song, W. Cai, M.-Z. Lee, Y. Zhou, H. Huang, S. Shan, M. J. Fulham, and D. D. Feng, "Lung nodule classification with multilevel patch-based context analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1155–1166, 2013.

[9] S. L. A. Lee, A. Z. Kouzani, and E. J. Hu, "Random forest based lung nodule classification aided by clustering," *Computerized medical imaging and graphics*, vol. 34, no. 7, pp. 535–542, 2010.

[10] F. Tixier, C. C. Le Rest, M. Hatt, N. Albarghach, O. Pradier, J.-P. Metges, L. Corcos, and D. Visvikis, "Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer," *Journal of Nuclear Medicine*, vol. 52, no. 3, pp. 369–378, 2011.

[11] C.-L. Huang, M.-J. Lian, Y.-H. Wu, W.-M. Chen, and W.-T. Chiu, "Identification of human ovarian adenocarcinoma cells with cisplatin-resistance by feature extraction of gray level co-occurrence matrix using optical images," *Diagnostics*, vol. 10, no. 6, p. 389, 2020.

[12] R. T. Leijenaar, G. Nalbantov, S. Carvalho, W. J. Van Elmpt, E. G. Troost, R. Boellaard, H. J. Aerts, R. J. Gillies, and P. Lambin, "The effect of suv discretization in quantitative fdg-pet radiomics: the need for standardized methodology in tumor texture analysis," *Scientific reports*, vol. 5, no. 1, p. 11075, 2015.

[13] M. Pomeroy, H. Lu, P. J. Pickhardt, and Z. Liang, "Histogram-based adaptive gray level scaling for texture feature classification of colorectal polyps," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. SPIE, 2018, pp. 507–513.

[14] J. Tan, Y. Gao, Z. Liang, W. Cao, M. J. Pomeroy, Y. Huo, L. Li, M. A. Barish, A. F. Abbasi, and P. J. Pickhardt, "3d-glcm cnn: A 3-dimensional gray-level co-occurrence matrix-based cnn model for polyp classification via ct colonography," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 2013–2024, 2019.

[15] G. Torres, S. Baeza, C. Sanchez, I. Guasch, A. Rosell, and D. Gil, "An intelligent radiomic approach for lung cancer screening," *Applied Sciences*, vol. 12, no. 3, p. 1568, 2022.

[16] N. Beig, M. Khorrami, M. Alilou, P. Prasanna, N. Braman, M. Orooji, S. Rakshit, K. Bera, P. Rajiah, J. Ginsberg *et al.*, "Perinodular and intranodular radiomic features on lung ct images distinguish adenocarcinomas from granulomas," *Radiology*, vol. 290, no. 3, pp. 783–792, 2019.

[17] J. L. L. Calheiros, L. B. V. de Amorim, L. L. de Lima, A. F. de Lima Filho, J. R. Ferreira Júnior, and M. C. de Oliveira, "The effects of perinodular features on solid lung nodule classification," *Journal of Digital Imaging*, pp. 1–13, 2021.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[20] AIM-Harvard, "Pyradiomics: an open-source python package for the extraction of radiomics features from medical imaging." 2016, accessed March 29, 2023. [Online]. Available: https://pyradiomics.readthedocs.io/en/latest/

[21] W. Pomp, "Parfor: a package to mimic the use of parfor as done in matlab." 2022, accessed April 15, 2023. [Online]. Available: https://pypi.org/project/parfor/