Explainable DL models in lung cancer using CLIP

Joel Bautista Rodríguez

Resum— La interpretabilitat continua sent un repte clau en els sistemes d'aprenentatge profund per al diagnòstic de càncer de pulmó. Aquest treball presenta CLIPMedical, un model multimodal que projecta descriptors radiològics estructurats i característiques de TC en un espai latent compartit mitjançant aprenentatge contrastiu. El corpus integra estudis anotats de tres hospitals catalans — Can Ruti, del Mar i Mútua Terrassa — i cinc variables clíniques (forma i densitat del nòdul, infiltració, diferenciació cel·lular i necrosi). S'aplica validació creuada de 5 folds amb recuperació *top-1* per avaluar cada camp. Amb extractors ResNet152, el model aconsegueix l'F1 més alt en infiltració (> 0,4) i *recall* puntuals > 0,7 en forma de nòdul, superant la configuració ResNet18 inicial, on la mediana d'F1 era 0,35. Els resultats mostren que l'alineació imatge-text pot augmentar la confiança clínica, tot i que el desequilibri de classes i les etiquetes mancants limiten la precisió en *cdiff*. Es proposa afegir una capçalera *multi-task* i depurar valors nuls com a treball futur.

Paraules clau— aprenentatge profund explicable; CLIP; càncer de pulmó; descriptors radiològics; model multimodal; tomografia computada.

Abstract— Interpretability remains to restrict clinical application of deep-learning networks to lung cancer diagnosis. Here, we introduce CLIPMedical, a contrastive multimodal model, that projects structured radiological descriptors to CT-derived image features in a shared latent space. The study leverages an annotated multi-institutional dataset for five critical descriptors of Can Ruti, del Mar and Mútua Terrassa hospitals: nodule shape, nodule density, infiltration, cell differentiation (cdiff) and necrosis. A 5-fold cross-validation with top-1 retrieval is used to test each descriptor. Compared to a ResNet18 backbone, the ResNet152 variant raises median F1 for infiltration from 0.35 to > 0.40 and achieves nodule-shape recalls of over 0.70 in best folds. These findings demonstrate that language-informed visual representation can offer intuitive textual explanations with little loss of classification performance. Remaining challenges include class imbalance and missing-value noise, and future work will explore multi-task heads and data curation as solutions to them.

Index Terms— explainable deep learning, CLIP, lung cancer, radiological descriptors, multimodal model, computed tomography.

1 INTRODUCTION – WORK CONTEXT

Iung cancer detection using deep learning (DL) has ,shown remarkable advancements in recent years, leveraging powerful models that extract meaningful patterns from medical images. However, one of the major challenges in medical AI applications remains the lack of interpretability, as deep learning models often operate as *black boxes*, making it difficult for clinicians to trust and validate their decisions. To address this, explainable deep learning (XDL) approaches are essential to bridge the gap between AI-generated predictions and medical reasoning.

This project focuses on leveraging CLIP to enhance explainability in lung cancer detection models. CLIP is a multimodal model capable of learning a shared latent space where images and textual descriptions are aligned**[8]**. While traditional deep learning models rely solely on image-based features, CLIP introduces a language-guided approach, enabling the association of lung cancer imaging data with medical reports. This allows for a more intuitive and interpretable decision-making process, aligning with the need for transparency in medical AI applications.

To achieve this, the project builds upon an existing lung cancer detection model that has already been trained to extract the most relevant visual features from radiological images. The primary goal is to extend this by effectively extracting and structuring features from medical reports, ensuring that textual and visual embeddings are properly aligned within a shared latent space. This is crucial for making CLIP fully functional in this specific medical application, as it ensures that the model learns the correct relationships between medical images and their corresponding diagnostic descriptions.

2 STATE OF THE ART

2.1 Medical Image Captioning and Report Generation

Nowadays, the automatic generation of medical reports from radiological images has emerged as a crucial area in medical AI **[1][2]**, with the purpose to bridge the gap between computer vision and natural language processing (NLP). The main goal is to translate visual medical data into accurate textual descriptions, enhancing diagnostic efficiency and reducing the workload of radiologists and clinicians. Retrieval-based and template-based approaches have been traditionally used but, in the last years, gradually replaced by deep learning (DL) - based generative models**[5][9]**, which allow for more flexible context-aware, and detailed medical report generation.

Thus, delving into deep learning-based methods for medical image captioning, modern approaches employ deep learning architectures to automate diagnostic report generation. These models generally adhere to an encoder-decoder paradigm**[9]**, categorized into three main types:

- CNN-Based Encoders: these use Convolutional Neural Networks (CNNs) like ResNet, EfficientNet, and Vision Transformers (ViTs) to extract meaningful visual features from medical images.
- RNN-Based Decoders: recurrent models such as RNNs, LSTMs, and Transformer-based architectures process these visual features to generate coherent textual descriptions.
- Attention Mechanisms: these enhance caption relevance by directing focus to the most critical image regions during text generation**[3].**

As mentioned, explainability is crucial actually because deep learning models must justify their decisions to gain tust from healthcare professionals. This is where CLIP and similar multimodal approaches play a vital role.

2.2 CLIP: A Game Changer in Medical AI

CLIP, developed by OpenAI, is a vision-language model designed to understand both images and text by jointly training on image-text pairs. CLIP is pre-trained on vast amounts of image-text pairs from the internet using contrastive learning**[3]**.

The way it works can be simplified to 3 steps:

1. Feature Extraction: CLIP encodes images and text separately using a VIsion Transformer (ViT) for images and a Transformer-based text encoder.

2. Latent Space Alignment: Both encoders project their outputs into a shared embedding space, where semantically related images and texts are pulled closer together**[4]**.

3. Zero-Shot Learning: Unlike traditional models that require fine-tuning on domain-specific datasets, CLIP is capable to classify images and generate text descriptions without retraining, making it highly adaptable to medical imaging tasks such as lung cancer detection, hematology and pathology (PLIP and BiomedCLIP), radiology report generation.

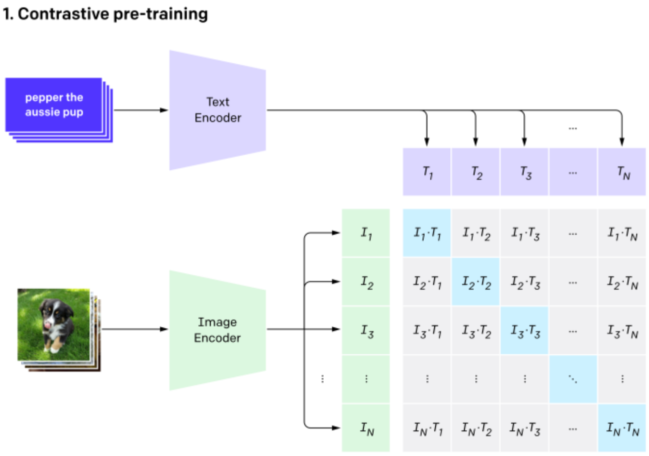

1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$

Figure 1. Contrastive pre-training

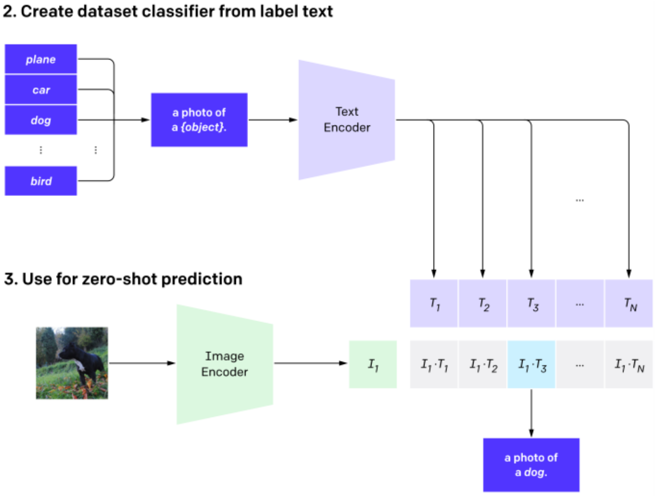

2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
T_1 T_2 T_3 ... T_N
I_1
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
a photo of a dog.

Figure 2. Dataset classifier from label text and zero-shot prediction

3 Objectives

The primary objective of this project is to enhance the explainability and interpretability of deep learning models for lung cancer detection by integrating medical image and diagnostic text embeddings into a shared latent space using CLIP-based architectures. This approach aims to create a more transparent and clinically relevant AI system that effectively aligns radiological images with textual medical diagnoses through contrastive learning.

A critical aspect of the project is modifying or adapting an existing CLIP architecture to integrate both the image encoder and the text encoder into a joint embedding space. Through contrastive learning, the model will be trained to maximize the similarity between images and their corresponding medical reports while increasing the distance between unrelated image-text pairs. To optimize this process, the contrastive loss function will be fine-tuned, ensuring that the alignment between visual and textual modalities is precise and meaningful in a medical context.

4 Metodology and planification

The project followed an *Agile methodology*, specifically applying the *Kanban approach* to manage workflow efficiently. For this purpose, the tool **Trello** was used to plan, organize, and monitor the progress of tasks throughout the project. This approach allowed for a clear visualization of task status, helped prioritize activities, and supported a flexible and iterative development process.

5 Data exploration

The data used comprises both radiological and clinical metadata extracted from annotated CT scans. On one hand, it includes pre-extracted image features that capture the texture of the CT scans on a per-slice basis. For each slice there is a feature vector and metadata, as well as an additional vector capturing intensity-related information. On the other hand, the dataset contains structured metadata which contains patient specific radiological descriptors.

In fact, for the training and evaluating the models, a subset of five clinically relevant variables was selected from this metadata: *nodule shape*, *nodule density*, *vinfiltration*, *cdiff* (cell differentiation) and *necrosis*.

5.1 Statistical Study of Radiologic Descriptors

To explore the structure and dependencies within the selected variables, variety of statistical indicators and visualizations were employed.

As a first step, the distribution of each selected variable was analyzed. Overall, the categorical variables displayed a clear imbalance across their classes. Among them, *vinfiltration* showed the most pronounced imbalance (see Figure 3). Otherwise, *nodule shape* showed the most balanced distribution as it can be observed in Figure 4.

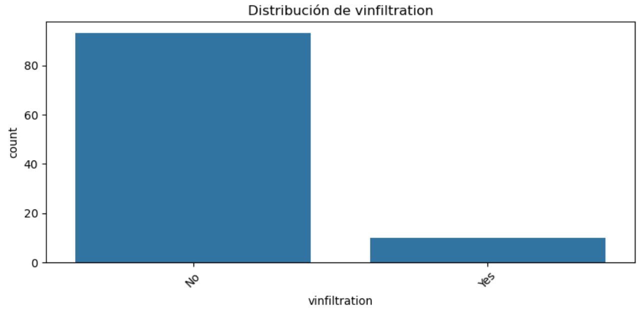


Figure 3. VInfiltration label distribution

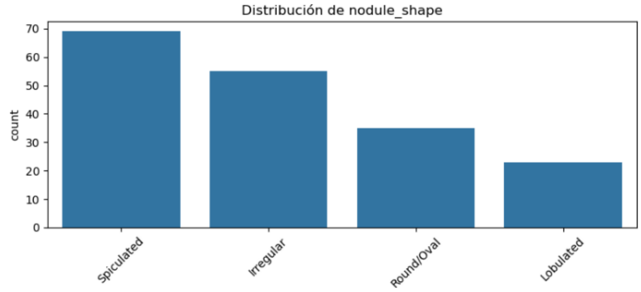


Figure 4. Nodule Shape label distribution

This imbalance is important to consider during model training and evaluation, as it may influence the classifier's sensitivity to minority classes.

In order to explore potential dependencies between variables, both *Chi-square* statistics and *Cramér's V* correlations were calculated across the descriptors. As shown in Figure 5, the χ^2 heatmap highlights notably strong associations between nodule shape and cdiff ($\chi^2 = 11$), as well as between nodule density and cdiff ($\chi^2 = 22$). This suggests that non-random associations may reflect underlying clinical or radiological patterns.

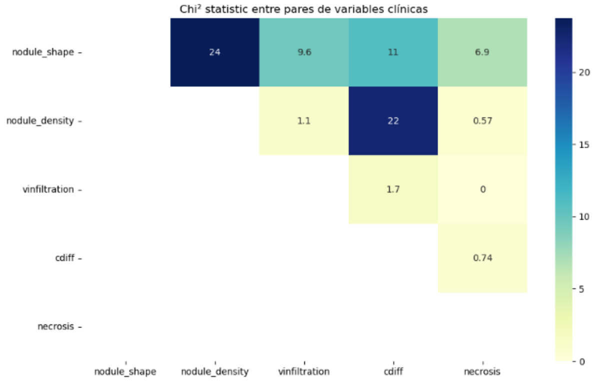


Figure 5. Chi-square heatmap

These findings are supported by the *Cramer's V* correlation matrix (Figure 6), where the strongest correlations also are between nodule density and cdiff (0.30).

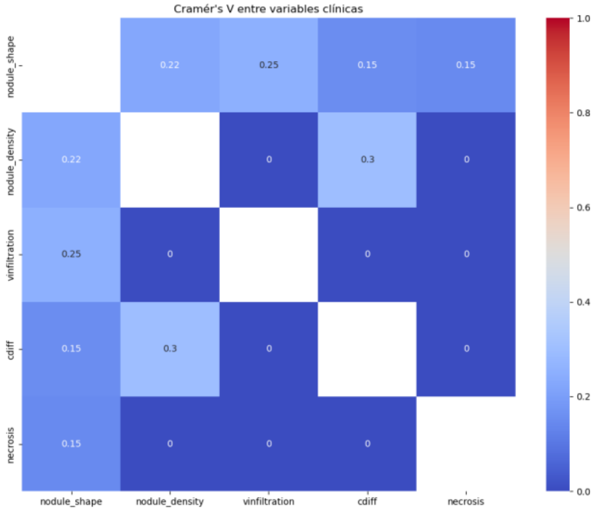


Figure 6. Cramer's V correlation matrix

6 Development

6.1 Architecture

6.1.1 CLIPMedical

To address the task of learning meaningful representations from both radiological image features and structured metadata, a multimodal neural architecture referred to as *CLIPMedical*, was implemented.

First, the clinical metadata is processed using a frozen ClinicalBERT transformer. Then, token embeddings extracted from ClinicalBERT are projected to a lower-dimensional space (from 768 to 512 dimensions) via learnable linear layer. Next, a multi-head self-attention mechanism is applied to model dependencies between the tokens, theorically producing a contextualized representation of the input descriptors

On the other side, image features are represented as 1024-dimensional vectors and passed through a feedforward projection layer consisting of a linear transformation, ReLU activation, and layer normalization. This has the objective to map the image embeddings into the same 512-dimensional latent space as the text embeddings.

The fusion of modalities is performed using contrastive alignment, where image and text embeddings are normalized and compared using Euclidean distance. The model outputs a similarity matrix between batch-wise image and text representations, promoting alignment of matched image-text pairs (see Figure 7).

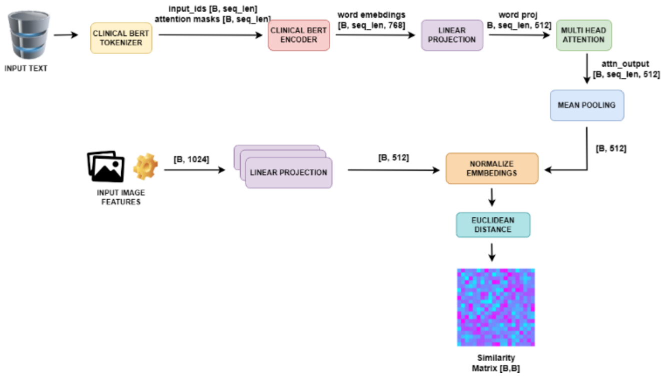


Figure 7. Architecture Schema

6.1.1 CLIPMedical With Attention in Image Features

CLIPMedical-Attn is a two-branch architecture that aligns clinical in-structured text and radiological slices into a common 512-dimensional latent space, integrating self-attention in both modalitites with contrastive loss.

Clinical metadata are tokenized and passed to a frozen ClinicalBERT encoder, which yields contextual token embeddings of dimension 768. The vectors are projected down to 512 dimensions via a learnable linear layer before being further optimized by a multi-head self-attention layer that models dependencies between descriptors. The branch outputs an l2-normalized sequence of 512-dimensional token vectors that represent the structured information of the patient.

For each patient, up to the twelve CT slices are encoded as 2048-dimensional feature vectors that incorporate GLCM and intensity information. One vector for each slice is created with mean pooling over channels. Each is projected from 512 dimensions through a linear layer. A learnable CLS token is prepended, and the sequence is passed through a light two-layer Transformer encoder-Normalised CLS embedding is utilized as the single 512-d image representation for the study.

Then, text and image embeddings are l2-normalized to make cosine similarity the same as Euclidean distance. Their pair-wise distances form a similarity matrix on which an InfoNCE loss is calculated. Training thereby pulls matched image-text pairs together and pushed away mismatches from each other, thus bringing together both modalities in the shared latent space.

Because the model keeps its image encoder, it is susceptible to contrastive pre-training in huge unlabelled sets and straightforward fine-tuning for downstream applications. Light heads for classification can be appended to frozen 512-dimensional embeddings, enabling retrieval, tagging or diagnostic prediction without the encoders being retrained and with compact parameter footprint.

6.2 Model Training

The training phase of the model is based on a contrastive learning approach, where the goal is to align clinical text descriptors with their corresponding image features in a shared embedding space. In addition, the model is trained in batches using paired samples, and optimization is guided by a contrastive *cross-entropy loss* that is explained below.

Each training batch starts with the construction of paried inputs. A batch of image feature vectors, each of image feature vectors is passed through a learnable projection head which results in a set of projected image embeddings of shape [B,512]. In parallel, the corresponding radiological descriptions already tokenized are processed by a frozen ClinicalBert encoder producing embeddings that are then pooled to obtain a final text representation of shape [B, 512].

All in a row, both the image and text embeddings are normalized, mapping them onto a unit hyperspace. This normalization step ensures that both cosine similarity and Euclidean distance can be interpreted in a comparable way for contrastive optimization.

Later, a pairwise distance matrix is computed across all image-text pairs in the batch. Specifically, the Euclidean distance is calculated between ecach image embedding and text embedding, resulting in a logits matrix of shape [B,B]. In this matrix, the entry at position *logits[i][j]* corresponds to the distance between the *i-th* image and the *j-th* text. These distances are later negated and scaled by a learnable temperature parameter.

The target labels for this training step are derived from the assumption that the *i.th* image in the batch corresoinds exactly to the *i-th* text. With the logits matrix and corresponding labels in hand, the model is trained using a *cross-entropy loss*, which encourages the similarity score of matching pairs to be maximized while penalizing similarity with the wrong pairs.

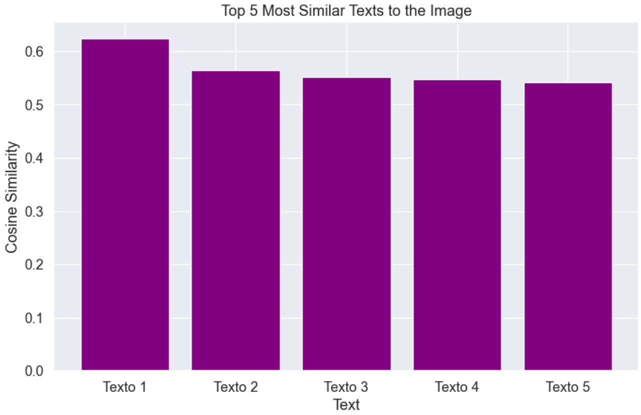
Finally, gradients are computed via backpropagation, and model parameters are updated using the Adam optimizer. This training loop explained above is repeated for each bach across multiple epochs.

6.3 Test & Evaluation

Once the model has been trained on a given fold, the evaluation phase begins by encoding and comparing test samples against a reference "catalog" constructed from the training data. This catalog serves as a fixed repository of text embeddings, allowing the model to retrieve the most semantically similar clinical descriptions for each unseen image.

To build this catalog, all text samples from the training set are decoded, tokenized and passed through the model's text encoder. A single 512-dimensional vector is obtained for each text, forming a matrix of embeddings that represents the reference catalog.

For inference, the model processes each test image feature vector which is projected into a normalized 512-dimensional embedding. Then, this embedding is compared to all entries in the catalog using cosine similarity, producing a ranked list of the most similar training texts (Figure X). From the list, only the top-1 prediction is used for evaluation metrics.



First, a qualitative evaluation is used to check whether the correct text is present among the top-k predictions for each patient. Second, each predicted text is parsed to extract its radiological descriptors which are also compared against the ground-truth descriptors for that sample. Then, a confusion matrix is generated, showing classification performance across categories (see Figure 9).

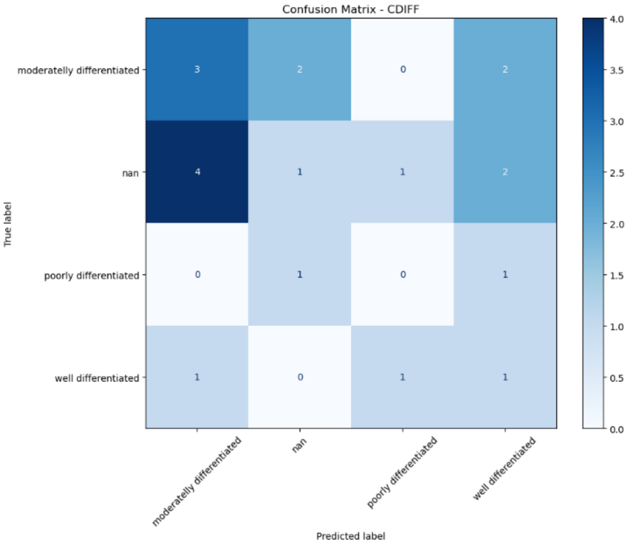


Figure 9. CDiff Confussion Matrix

Finally, the quantitative performance is measured using standard classification metrics. Precision, recall, f1-score and accuracy are computed for each radiological descriptor on the top-1 predicted text. These metrics are computed for each fold and then aggregated across folds to obtain an overall performance summary. All results are saved to CSV files and plotted using boxplots, as seen in Figure 10.

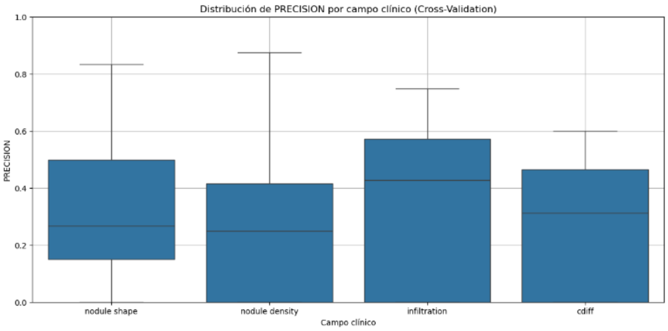


Figure 10. Precision for each descriptor across all folds

7 Experimental Design

This experimental setup is designed to evaluate the performance and generalizability of the *CLIPMedical* in the task of aligning clinical text descriptors and imatge-derived features.

To this end, as explained in the data exploration section, a multi-hospital data set was used, combining annotated cases from Can Ruti, Del Mar, and Mutua Terrassa. For all the experiments, the same five radiological descriptors explained before were selected.

To ensure robust evaluation, the experimental protocol followed a 5-fold cross validation strategy. For each fold, a new model instance was trained from scratch.

Stratification was not enforced, but folds were randomly sampled to reduce potential class bias.

Each experiments has itself a set of paràmetres, which are selected to explore how the model performs with the changes in any of them. These parameters include the batch size, number of epochs, pre-trained extractor used, learning rate and optimizer, among others. Before presenting each experiment, the specific configuration used will be explicitly stated to provide clarity.

8 Pre-Trained Extractors Comparative

The configuration of parameters used for training and evaluation in the following experiments is summarized in Table 1.

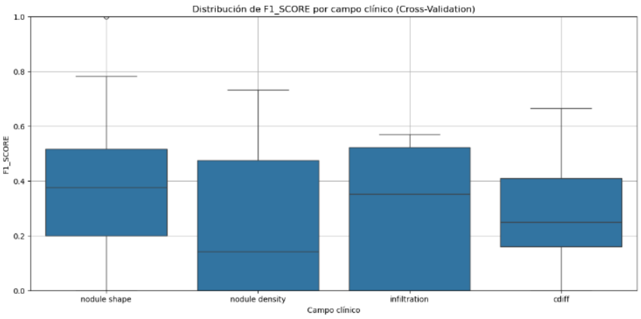
Parameter	*Value*
Batch size	64
Epochs	30
Learning rate	$1x10^{-4}$
Optimizer	Adam
Loss function	Cross-Entropy Loss
Tokenizer	ClinicalBERT
Text Encoder output	768 → 512
Image Encoder output	1024 → 512
Evaluation Strategy	Top-1 retrieval + per field metrics

Table 1. Parameters Configuration

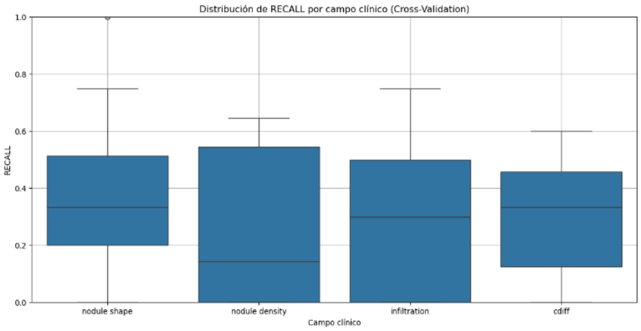
8.1 ResNet18

The evaluation of the model using imatge features extracted from a pretrained ResNet18 revealed generally poor performance. The overall predictive capacity was low, and results varied considerably across clinical fields.

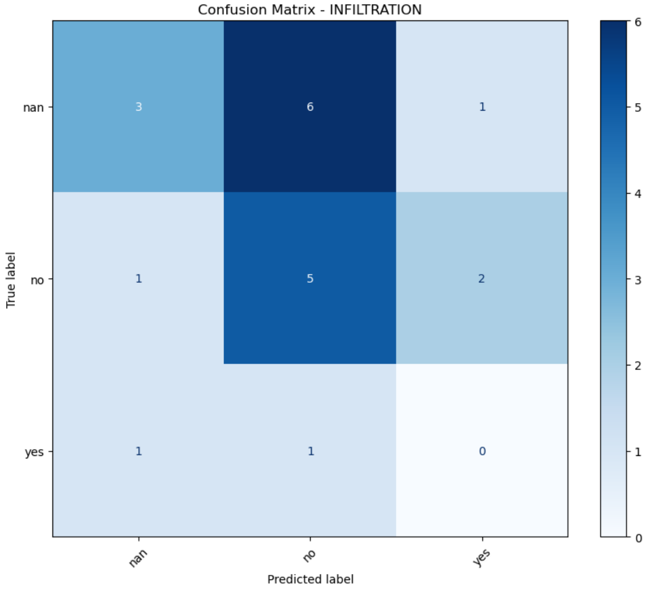
Globally, infiltration was the descriptor with te highest stability and overall F1-score, showing a median above 0.35 but without compact interquartile range. On the other hand, nodule density and cdiff showed more dispersion and lower median vàlues, indicating inconsistent prediction quality across folds. This can be promoted in part by class imbalance, especially in the case of cdiff, where "poorly differentiated " yielded very low precision (0.10) and recalll (0.17), as shown in figures 11, 12 and 13.



From a recall perspective (Figure 11), nodule shape and infiltration reached the broadest range, with some folds achieving values above 0.70. However, these high scores are not sustained consistently, as reflected in the wider spred of the boxplots.



When looking at the overall F1-score distribution (Figure 13), infiltration and nodule shape appear to be the most learnable and consistently predictable fields, while cdiff continues to reflect the impact of low recall in minority classes. In particular, the standard deviation for "well differentiated" in cdiff is high, indicating inconsistent behavior across folds.

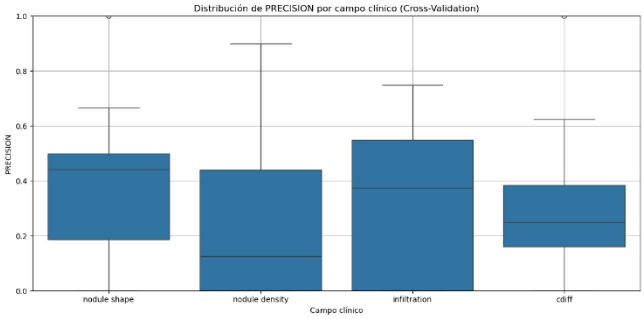


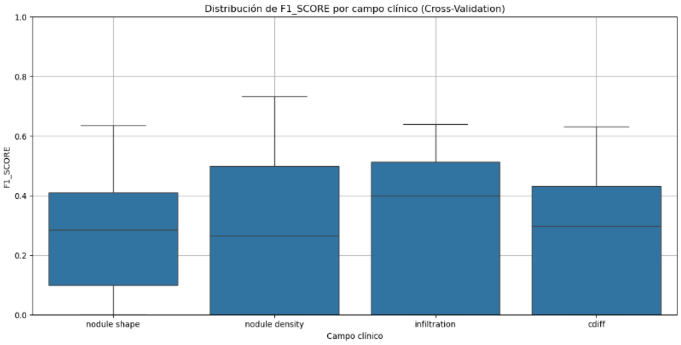
8.2 ResNet152

Evaluation of ResNet152 resulted in slightly improved yet still limited performance compared to the ResNet18 setup. While overall results remain away from what is being sought, there is a noticeable increase in consistency across folds for several clinical descriptors, particularly in terms of precision.

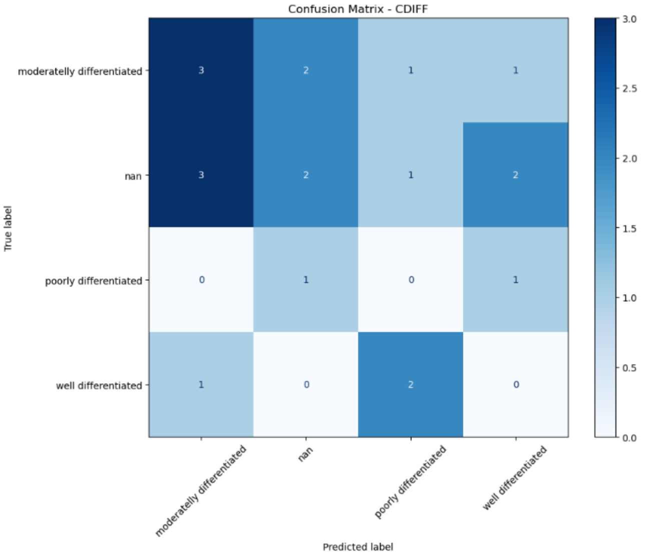
The *infiltration* descriptor continues to stand out as one of the most consistently predicted fields, achieving the highest average F1-score and balanced Distribution across folds (see Figure 14). However, the confusion matrix (Figure 15) reveals that the model still tends to confuse "yes" and "no" cases with missing values (nan), indicating room for improvement in data quality or a new strategy that should mean taking out this missing values.

In contrast, *cdiff* remains one of the most challenging fields for the model. Despite a slight increase in F1-score for the "moderately differentiated" class, the model fails to consistently predict "poorly" and "well diferentiated" samples as seen in Figure 16.





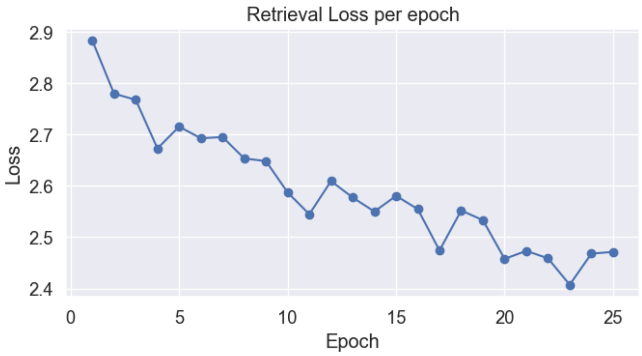
Precisions scores were generally higher and more estable than in ResNet18, indicating an improved ability to reduce false positives. *Infiltration*, in particular, reached the highest median precision across all descriptors, as seen in Figure 17 which suggests that the model was more confident and accurate when predicting this class.



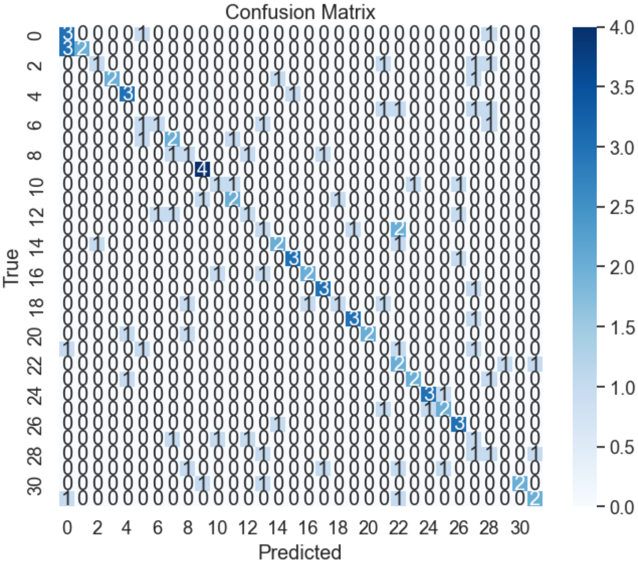
On the other hand, recall remained more irregular (Figure 18), especially for classes with fewer samples, such as "poorly differentiated" in the *cdiff* field. The model often failed to retrievethose underrepresented classes, leading despite improvements in precision.

8.3 MobileNetV2

The retrieval-loss curve indicates a healthy training process: after an initial sharp drop from ~2.9 to 2.75 in the first few epochs, when the model learns the most basic retrieval patterns, the drop becomes more stable, falling in small, jagged steps rather than a straight line. Those minor oscillations (i.e., the bump at epochs 11–12 and 17–18) are to be expected in stochastic optimization and suggest that the learning rate and batch variability introduce some noise without interfering with convergence. Past epoch 10 the loss flattens out at around 2.6, then continues again on a more apparent downwards trend, hitting a low of about 2.45 by epoch 23 before flattening out. In general, the consistent decline of around 0.45 points for each of the 25 epochs indicates that the model is continuing to refine its retrieval representations.

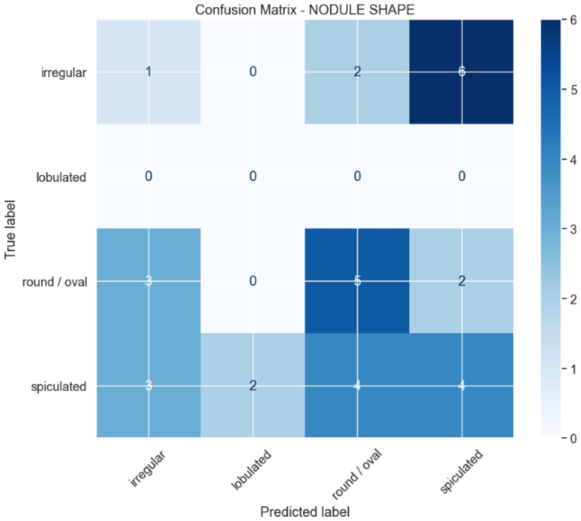


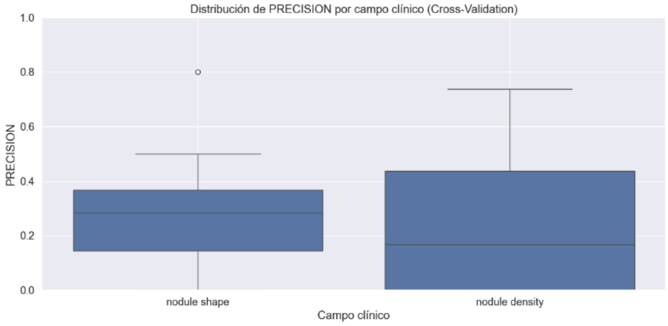
In the following cross-modal confusion matrix (Figure X) the text embeddings are rows (radiology reports) and the image embeddings are columns (the corresponding CT scans) for all 32 patients in each batch. A cell signifies how often a text query's closest neighbor in the common space is an image of a particular patient; thus, a main-diagonal value signifies that patient i's text has correctly matched to the image embedding of the patient i. The diagonal is almost entirely filled out with three-to-four hits for every patient and off-diagonal cells are mostly zero and only sporadically register one mismatch. This implies that, on the training set, the model receives the correct patient's image for almost any text query—attaining near-perfect text-to-image correspondence at the patient level. The extremely small number of solitary errors are not clumped together between specific pairs of patients, suggesting they are the consequence of the very small sample size per patient and not from a systematic confusion. Generally, the matrix suggests excellent intra-patient retrieval performance on training data, but the very high accuracy also advises caution that this behavior generalizes to unseen patients in a test or validation split to rule out overfitting.



Evaluation of MobileNetV2 yielded modest yet highly variable results in comparison with the lightweight ResNet18 baseline. Although the architecture remains attractive for its computational efficiency, overall scores still fall short of the desired threshold; moreover, dispersion between folds is pronounced for most clinical descriptors, as reflected by the wide inter-quartile ranges in precision (see Figure X).

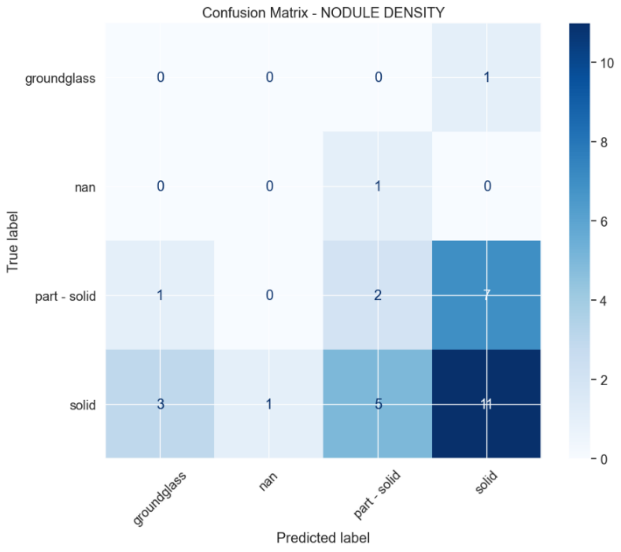
The confusion matrix (see Figure X) draws out that the classifier performs poorly in distinguishing nodule shapes, especially between irregular and spiculated ones. Out of the nine nodules which are actually irregular, only one is correctly classified, while six are mis-classified as spiculated and two as round/oval. The spiculated class also does comparatively poorly, only four out of thirteen cases are classified accurately, the rest being spread out across all the other classes. Overall, approximately one-third of the 32 cases fall on the diagonal, indicating that the model confuses spiculation pointed extensions with jagged irregular edges and has not got representative samples for lobulated nodules.





The nodule-density confusion matrix (see Figure X) points out how extreme class imbalance deflates the model's discrimination capability. Solid nodules dominate the training split (≈20 samples of 32). Faced with such skewed data, the network is skewed towards the majority class: all ground-glass nodules are incorrectly classified as solid, one nan case is being assumed as part-solid, and seven of the ten part-solid nodules are also pulled into the solid class. Even within the most represented class, there is less than optimal performance (eleven correct and nine errors showing that over-representation alone does not always lead to precise learning.

The trend shows that the model has learned a generic “solid-like” signature but not enough signal to detect the more nuanced differences in attenuation that distinguish ground-glass and part-solid lesions. Preventing this will require rebalancing techniques and possibly more telling radiomic features so that the classifier is not nearly solely based upon the number of solid examples.



9 Conclusion

This research set out to see whether a contrastive, language-enabled representation could bridge the interpretability gap in lung-cancer imaging. While the cross-modal approach was able to register CT images with structured text descriptors on technical grounds, the empirical evidence collected here is not yet adequate to demonstrate that the procedure is clinically viable.

Retrieval loss decreased step by step at the first glance while training confusion matrices registered nearly perfect patient-level matchings; however, the same matrices when used for validation data showed extreme vulnerability in terms of deteriorating drastically whenever the class distributions skewed or the descriptors were sparsely labeled. For example, all ground-glass nodules got misclassified as solid, while minority classes of cell differentiation were almost not observed by the model. These failures imply that the system is learning dominant trends and not acquiring semantically meaningful boundaries — precisely the opposite of its role as an explainable system.

Methodologically, stronger backbones such as ResNet152 did improve brute metrics, but even the best configuration plateaued far below clinically useful recall for under-represented findings. The experiment thus exposes a bitter truth: contrastive alignment alone cannot beat poor or skewed clinical data. In the absence of extreme balancing methods, richer annotation, and outside validation, the promise of transparent, text-anchored decision support remains an illusion.

In summary, the work presents a cautionary proof-of-concept: language-informed visual embeddings would make lung-cancer evaluation more transparent, but the work here is not successful in making the concept viable in its current form. Substantial methodological advancement and more detailed, better-balanced datasets are required before such a system would be embraced by radiologists or change standard clinical practice.

10 Future improvements and next steps

To drive the concept from proof-of-principle to clinical usefulness, future studies need to prioritize shoring up the foundations rather than merely increasing the size of the current pipeline. The priorities are as follows:

1. Balanced, enriched data – Use systematic over-sampling, synthetic augmentation, or curriculum learning so that minority phenotypes (ground-glass or poorly differentiated lesions, for instance) yield a strong training signal.

2. Multi-task heads – Replace single-label classifiers with architectures that model correlated descriptors jointly (shape, density), allowing the network to lend statistical strength across related tasks.

3. Semi-supervised pre-training – Leverage large repositories of unlabeled CT scans to learn strong visual representations before fine-tuning on limited, expertly annotated subset.

These three aspects must be addressed to transform the current, imbalance-sensitive prototype into a trustworthy decision-support tool.

Agraïments

...., ...

Bibliografia

[1] Referència 1
[2] Referència 2
[3] Etc.

APÈNDIX

A1. SECCIÓ D'APÈNDIX

....

A2. SECCIÓ D'APÈNDIX

....