# 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 dos tipus d'aprenentatges. 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-k* per avaluar cada variable. Amb l'extractor ResNet152, el model aconsegueix l'*F1* més alt i *recalls* puntuals de aproximadament 0,75, superant la configuració ResNet18 inicial, on la mitjana 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. A més s'exploren dos mecanismes d'aprenentatge i dos mecanismes d'agregació de característiques diferents.

**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 corpus integrates annotated studies from three Catalan hospitals: Can Ruti, del Mar, and Mútua Terrassa, along with five clinical variables (nodule shape and density, infiltration, cell differentiation, and necrosis). A 5-fold cross-validation with top-k retrieval is applied to evaluate each variable. With the ResNet152 extractor, the model achieves the highest F1 score and recall values of approximately 0.75, surpassing the initial ResNet18 configuration, where the average F1 was 0.35. The results demonstrate that image-text alignment can increase clinical confidence, although class imbalance and missing labels limit accuracy. In addition, two different learning approaches and two aggregation mechanisms of visual features have been explored and tested.

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

## 1 Introduction – Work Context

Lung cancer detection using deep learning (DL) has shown remarkable advancements in recent years [1][2], 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 [3][4], 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 [4] to bridge the gap between AI-generated predictions and medical reasoning.

The aim of this project is to use CLIP to enhance explainability in lung cancer detection models. CLIP is a multimodal model capable of learning a joint latent space [5] where text descriptions and images are aligned. Compared to traditional deep learning models which would make use of only image-based features, CLIP suggests a language guided approach [6], enabling mapping of lung cancer imaging data onto medical reports. This allows for a more understandable and intuitive decision-making process that can correspond to the need for transparency in medical AI usage.

This is achieved by basing the project on a pre-existing lung cancer detection model that has already received training in order to extract the most beneficial visual features from radiological images. The focus is to improve this through successfully capturing and structuring features from medical reports while making textual and visual embeddings consistent within a shared latent space [7]. This is important in order to get CLIP to be fully operational in this particular medical use case, as it guarantees that the model learns the right relationships between medical images and their respective diagnostic descriptions.

Currently, lung cancer diagnosis relies heavily on computed tomography (CT) scans, which are three-dimensional imaging studies used to visualize the internal structures of the lungs in high detail. Radiologists interpret these CT scans by analyzing specific visual patterns and characteristics that are indicative of disease. Each scan is typically associated with a set of structured radiological descriptors that inform the clinical assessment and treatment planning.

## 2 STATE OF THE ART

### 2.1 Medical Image Captioning and Report Generation

Nowadays, computer aided reporting of medical reports from radiological images is a significant area of medical AI [7][8], which tries to bridge the gap between computer vision and NLP. The underlying goal is to transform visual medical information into accurate textual information for enhanced diagnosis efficiency and workload of radiologists and clinicians. Retrieval-based as well as template-based approaches have been used conventionally but, more recently, sequentially replaced by deep learning (DL)-based models [9], enabling context-aware, more flexible, and elaborated generation of medical reports

Thus, delving into deep learning-based medical image captioning, current methods leverage deep learning architectures for diagnostic report generation automation. These models generally adhere to an encoder-decoder architecture, which is divided into three general categories [11]:

- CNN-Based Encoders: employ Convolutional Neural Networks (CNNs) like ResNet, EfficientNet, and Vision Transformers (ViTs) in extracting useful visual features from medical images.
- RNN-Based Decoders: recurrent models such as RNNs, LSTMs, and Transformer-based models decode these visual features to generate coherent text descriptions.
- Attention Mechanisms: these enhance caption relevance by prioritizing the most relevant image regions during text generation [10].

As mentioned, explainability is crucial because deep learning models must explain their decisions in order to receive trust from healthcare experts [4]. This is where CLIP and other multimodal approaches have a crucial role to play.

### 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 [1]. CLIP is pre-trained on vast amounts of image-text pairs from the internet using contrastive learning.

The way it works can be simplified to 3 steps:

1. Feature Extraction: CLIP encodes images and text separately using a Vision Transformer (ViT) [12] for images and a Transformer-based text encoder.
2. Latent Space Alignment: Both encoders project their outputs into a shared embedding space [1] [6], where semantically related images and texts are pulled closer together.
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 [1], making it highly adaptable to medical imaging tasks such as lung cancer detection, hematology and pathology (PLIP and BiomedCLIP), radiology report generation [13][14].

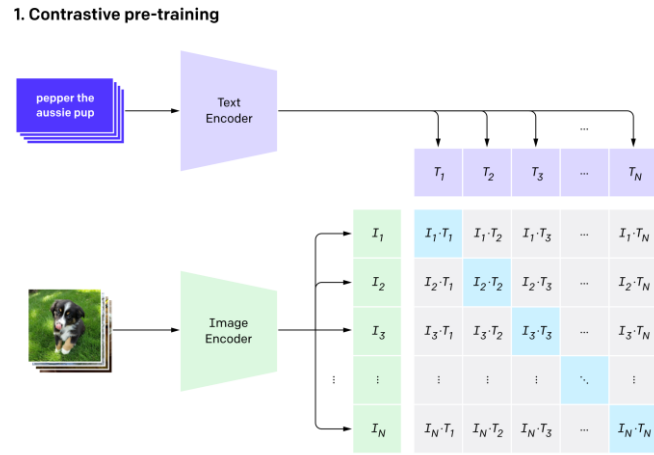


*Figure 1. Contrastive pre-training*

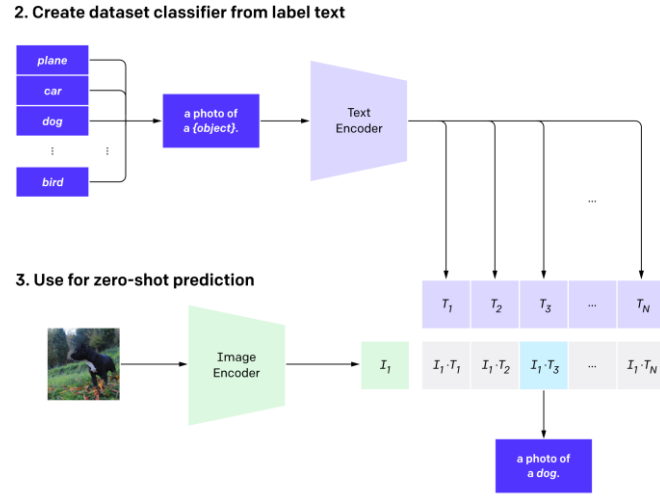


*Figure 2. Dataset classifier from label text and zero-shot prediction*

## 3 OBJECTIVES

The main objectives of the project are centered on scaling explainability in lung cancer detection with multimodal modeling. Specifically, the challenge is to integrate pre-trained visual features from images into structured radiological descriptors in a single representation space.

The objectives of the project are:

1. Create a CLIP-based model that can map pre-trained visual features learned from CT scans in-

to structured radiological descriptors to enable interpretable image-text correspondences for medical data.
2. Explore visual feature aggregation mechanisms, in particular testing the appropriateness of mean pooling as a baseline and the application of an attention-based mechanism for selection and prioritization of the most diagnostic features from multiple CT slices.
3. Use and compare a suggested loss function, a CrossEntropy-based loss that is entirely described in the architecture and training sections. Another hard mining triplet loss will be tried to check its impact on embedding space quality.
4. Compare different visual representation spaces by looking at how different pretrained backbones (like ResNet18, ResNet152, MobileNetV2) affect the model's ability to align image and text modes.

## 4 METODOLOGY AND PLANIFICATION

The project followed an *Agile methodology*, specifically applying the *Kanban approach* to manage workflow efficiently [18]. For this purpose, as seen in Figure 20, the tool **Trello** was used to plan, organize, and monitor the progress of tasks throughout the project [19]. This approach allowed for a clear visualization of task status, helped prioritize activities, and supported a flexible and iterative development process.

## 5 DATASET

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.

The dataset used in this project is a multi-hospital one, where the annotated cases of lung cancer at Can Ruti, Hospital del Mar, and Mutua Terrassa are joined together.

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*.

Each of this measures capture distinct aspects of tumor characterization. Shape describes the external morphology of the lesion, perhaps providing a hint at its potential for malignancy. Density captures the internal composition of the nodule between its different appearances. Infiltration assesses whether the lesion invades adjacent lung structures, which is generally a marker for more advanced disease. Cell differentiation (cdiff) refers to the degree to which the tumor cells resemble normal tissue, with lower differentiation typically indicating more aggressiveness. Necrosis is the presence of dead tissue in the tumor, perhaps being a marker of rapid growth.

### 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, as showed in Table 8. Overall, the categorical variables displayed a clear imbalance across their classes [20]. Among them, *vinfiltration* showed the most pronounced imbalance (see Figure 4). Otherwise, *nodule shape* showed the most balanced distribution as it can be observed in Figure 3.

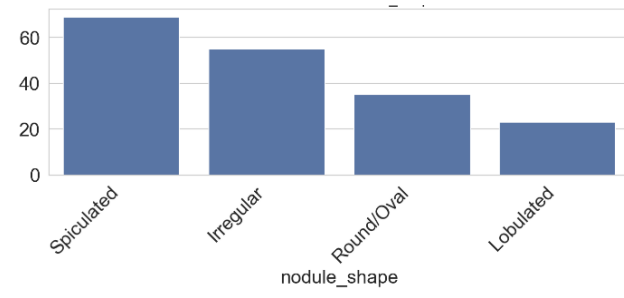


*Figure 3 Nodule Shape Histogram*

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

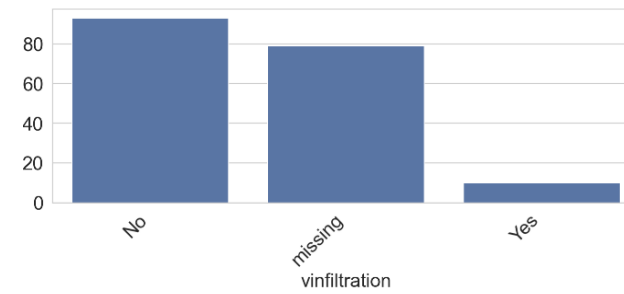


*Figure 4 vInfiltration Histogram*

In order to explore potential dependencies between variables, both *Chi-square* statistics and *Cramér's V* correlations were calculated across the descriptors [21]. The Chi-square ($\chi^2$) statistic is calculated on a contingency table between pairs of categorical variables. It measures the extent to which the observed frequencies of the categories deviate from the frequencies that would be expected if the variables were independent.

As shown in Figure 5, the $\chi^2$ 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.

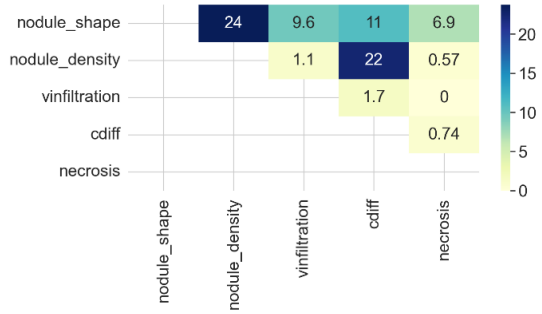


*Figure 5 Chi-square matrix of correlations*

These findings are supported by the *Cramer's V* correlation matrix (Figure 6) [22], where the strongest correlations also are between nodule density and cdiff (0.30).

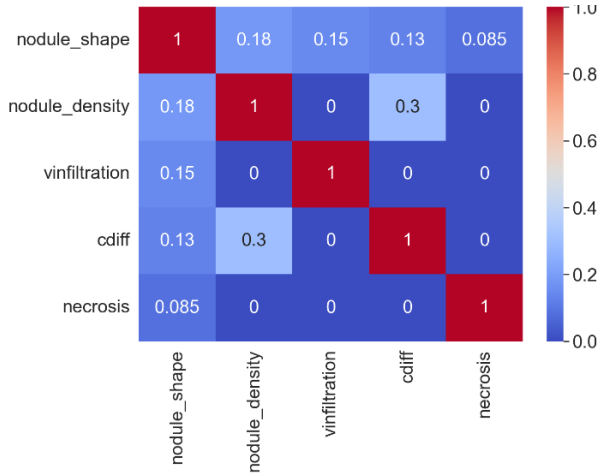


*Figure 6 Cramer's V correlation matrix*

This statistic is a normalized version of the *chi-square* statistic, scaled for total sample size, with a correction against overestimating the association in small or unbalanced tables. It is computed by first computing $\chi^2$, normalizing it for the total sample size, and then correcting for it. The result can range from 0 (no association) to 1 (complete association).

## 6 DEVELOPMENT

### 6.1 Architecture: 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 (the five descriptors together) is processed using a frozen ClinicalBERT transformer [23]. This decision of combining the descriptors texts into a single embedding was made independently. 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 all together.

On the image side, image features are initially extracted as 1024-dimensional vectors from individual slices of CT scans. The raw features are then projected by a feed-forward layer that involves a linear transformation, ReLU activation, and layer normalization. To aggregate slice-level information into a single patient-level representation, two methods are explored: a baseline method using average pooling over all available slices, and an attention over a constant subset of 15 centrally selected slices. In the latter, the model learns slice-level importance weights using a learnable attention module that calculates relevance scores using a small neural network followed by a softmax operator. This weighted average is integrated into the main training loop so the model can focus on the most informative slices to match images with clinical text.

The fusion of modalities is performed by aligning [16] image and text embeddings that are normalized and compared using cosine similarity. The model outputs a similarity matrix between batch-wise image and text representations, promoting alignment of matched image-text pairs (see Figure 7).

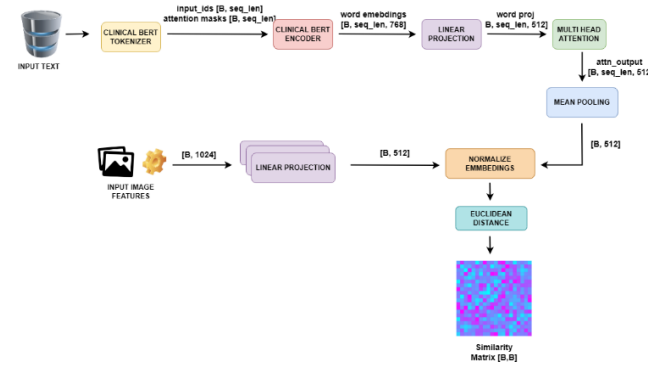


*Figure 7 Architecture Schema*

The training phase of the model is based on a supervised learning approach, where the goal is to align clinical text descriptors with their corresponding image features in a shared embedding space [1]. In addition, the model is trained in batches using paired samples, and optimization is guided by a *cross-entropy loss* [17] 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 [16], mapping them into a unit hyperspace. This normalization step ensures that cosine similarity can be interpreted in a comparable way for matching optimization.

Later, a pairwise distance matrix is computed across all image-text pairs in the batch. Specifically, the cosine similarity is calculated between each 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.

As explained in one of the objectives of this project, and as will be explored in one of the experiments, the same model is also trained using a contrastive loss method. This approach is based on a triplet loss mechanism, whereby the model is trained to distinguish between matching and non-matching image-text pairs by drawing the correct image-text pair closer and pushing the other pairs apart. To build this loss, we compute a logits matrix containing the cosine similarity between all image and text embeddings in the batch. We use the right corresponding text as a positive pair for all images in the batch and select other texts in the batch as negative pairs.

In hard negative mining, instead of randomly sampling negative pairs, we focus on selecting the ones that are closest to the positive pair, which generates harder triplets. As can be seen in Figure 8, such "hard negatives" are near the anchor point and are the most challenging examples for the model. The contrastive loss function punishes the model if the similarity of the positive pair is less than the similarity of a hard negative pair, in addition to some margin. The model is trained to align the image and text embeddings correctly in a shared space, also ensuring that the negative pairs do not resemble too closely the positive pair, as can be seen from the diagram where hard negatives are placed close to the anchor in the space. This forces the model to learn from the most difficult examples, thereby increasing its ability to correctly distinguish between matching and non-matching pairs.

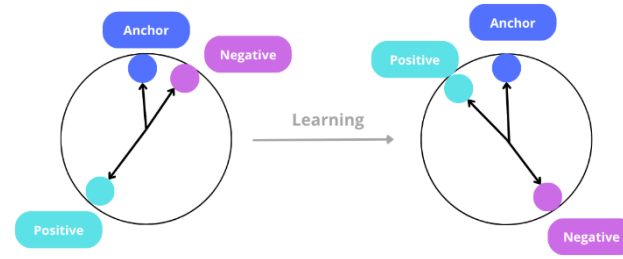


*Figure 8 Triplet Margin Loss Schema*

### 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. As seen in Figure 9, 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 [24].

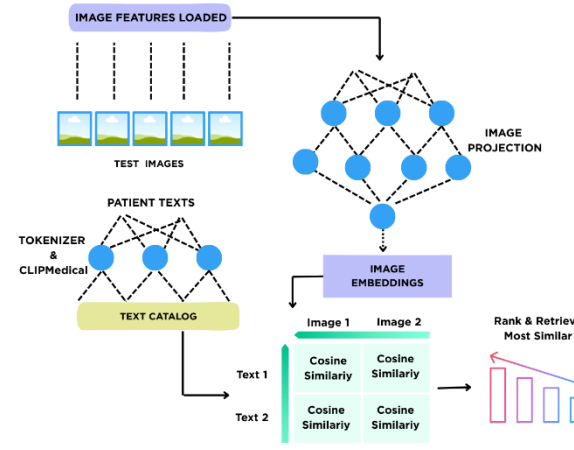


*Figure 9 Inference and evaluation schema*

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 catalog texts. From this list, instead of limiting the examination to the top-1 only, several values of k are considered (e.g., top-1, top-3, top-5) in order to verify for how often the correct clinical description appears in the top-k most similar candidates. This provides a more detailed view of the model's retrieval capability.

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 sam-

ple. Then, a confusion matrix is generated, showing classification performance across categories.

Finally, the quantitative performance is measured using standard classification metrics [25]. Recall, precision, F1-score and accuracy are computed for each radiological descriptor on the top-k 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.

## 7 Experimental Design

The goal of the experiments is to evaluate the performance and generalizability of the CLIPMedical model in the task of aligning clinical text descriptors with image-derived features. For this purpose, the following experiments were conducted to study the effects of different model settings and test modes:

- Experiment 1 – ***Visual feature extractor comparison***: this experiment explores the effect of the visual representation on model performance. Pre-trained backbones (ResNet18, ResNet152, and MobileNetV2) are compared to test which visual space is better for textual information.
- Experiment 2 – ***Top-k performance***: the experiment calculates how the performance of the model varies with considering different numbers of top-k candidates. Instead of relying entirely on top-1 predictions, additional testing is performed for top-3 and top-5 retrievals.
- Experiment 3 – ***Impact of contrastive learning***: here, the model is trained with the Triplet Margin Loss function, instead of the standard cross-entropy objective. The goal here is to investigate whether this contrastive loss improves the latent space structure and retrieval performance.
- Experiment 4 – ***Nodule representation space comparision***: in experiment 4, two methods for aggregating image features from CT scan slices are contrasted: an average pooling baseline and an attention-based approach. The aim is to evaluate which of these approaches yields more effective global representations of the images to perform alignment downstream.

To ensure robust evaluation, the experimental protocol followed a 5-fold cross validation strategy [26]. 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 experiment has itself a set of parameters, 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.

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

| *Parameter* | *Value* |
|---|---|
| *Batch size* | 32 |
| *Epochs* | 25 |
| *Learning rate* | $1x10^{-4}$ |
| *Optimizer* | Adam |
| *Loss function* | Cross-Entropy Loss or Triplet Margin Loss (Experiment 3) |
| *Tokenizer* | ClinicalBERT |
| *Text Encoder output* | 768 → 512 |
| *Image Encoder output* | 1024 → 512 |
| *Evaluation Strategy* | Top-k retrieval + per field metrics |

*Table 1 Parameters Configuration*

## 8 Results

### 8.1 Visual representation space comparision

To compare the impact of different visual extractors on the model performance, the following sections present comparison analysis for each configuration. For each of the three pre-trained backbones (ResNet18, ResNet152, and MobileNetV2), two tables are included: one for training and one for testing performance across all five clinical descriptors. Mean and standard deviation of recall, precision, and F1-score across all cross-validation folds are given in each of the tables. Additionally, a line plot is displayed to indicate the trend of training loss through epochs, presenting information on the model's behavior upon convergence under every setting.

#### 8.1.1 ResNet18

Figure 10 shows the retrieval loss over training epochs for the ResNet18 model. The graph is a smooth and steady downward line, from more than 3.2 at the start right down below 2.7 by epoch 25. The initial few epochs have a steep drop, showing good early learning, and later epochs have decreasing but steady improvements. There is no sign of overfitting or wild peaks and valleys. Overall, the trend shows that the model is converging well, but further epochs can still yield marginal increases.

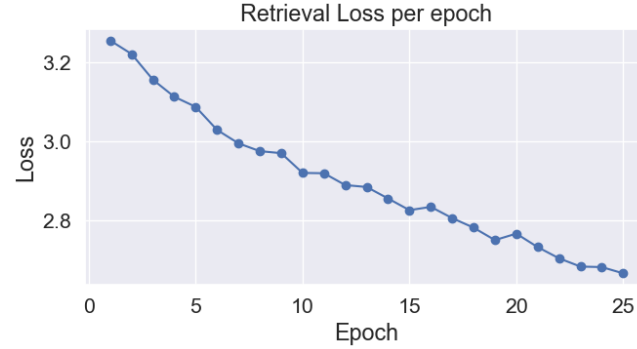


*Figure 10 ResNet18 Loss Curve*

On the training set (Table 2), the model performs

relatively stable across the board. F1-score is most prominent for necrosis (0.59 ± 0.00), then cdiff (0.55 ± 0.01), with nodule shape and nodule density slightly lower around 0.43. Precision is greater than recall in all descriptors, showing the model is biased towards making braver but more confident predictions.

| Test | c d i f f | infiltration | necrosis | nodule_density | nodule_shape |
|---|---|---|---|---|---|
| recall | 0.31±0.21 | 0.39±0.22 | 0.58±0.36 | 0.27±0.21 | 0.30 ± 0.13 |
| precision | 0.34±0.17 | 0.39±0.18 | 0.62±0.35 | 0.29±0.26 | 0.30 ± 0.09 |
| F1-score | 0.27±0.14 | 0.36±0.16 | 0.57±0.03 | 0.24±0.18 | 0.27± 0.10 |

*Table 3 Test Metrics for ResNet18*

But the performance is much worse on the test set (Table 3). Here the F1-scores reduce across all the variables, the most affected being cdiff (0.27 ± 0.14) and nodule density (0.24 ± 0.18), which also have large variance across folds. Necrosis, on the other hand, generalizes quite well (F1-score of 0.57 ± 0.03), indicative that it may be easier to train from the provided visual features.

Overall, these results suggest a seeming disparity between test and training performance, especially for descriptors like cdiff and nodule density, which are also known to be class-imbalanced. This shows that although the model can learn some structure from the training data, it does not generalize well, more likely due to insufficient diversity of samples and imbalanced distributions.

### 8.1.2 ResNet152

Figure 11 illustrates the los curve during training using the ResNet152 extractor. The line begins higher than 3.2 and shows a steep, consistent decline across all 25 training epochs, rising below 2.6 towards the end of training. Contrary to the case with ResNet18, the drop here is not only sharper in the initial epochs but also more smooth and prolonged with time and has less fluctuation and lower end-loss values. The trend shows that the model learns not only faster but more steadily as well. In general, this path shows improved convergence and fitting to the retrieval task compared with ResNet18.

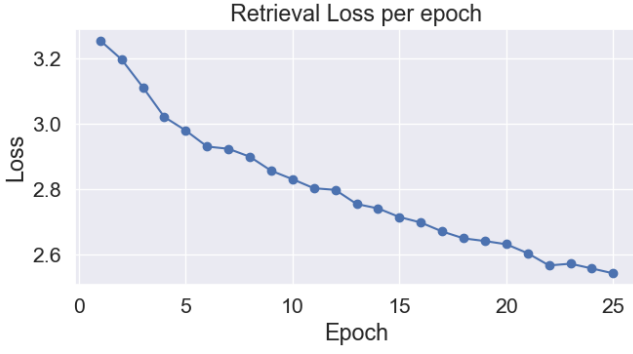


*Figure 11 ResNet152 Loss Curve*

In the cross-modal confusion matrix (Figure 12) 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 matching capability also advises caution that this behavior generalizes to unseen patients in a test or validation split to rule out overfitting.

The performance of the ResNet152 model is shown in Tables 4 (train) and 5 (test), and it has the most balanced and stable performance among the three extractors being tested.

On the training set (Table 4), ResNet152 achieves high performance on almost all descriptors. Necrosis again achieves the best with an F1-score of 0.68 ± 0.02, while infiltration (0.56 ± 0.02) and cdiff (0.48 ± 0.00) also achieve high performance. The nodule shape and nodule density achieve slightly lower but still consistent performance (0.42 and 0.46 respectively), with good generalization between descriptor types.

| Train | cdiff | infiltration | necrosis | nod-ule_density | nodule_shape |
|---|---|---|---|---|---|
| recall | 0.38±0.00 | 0.40±0.00 | 0.61± 0.13 | 0.28±0.04 | 0.27±0.01 |
| precision | 0.26±0.04 | 0.31±0.02 | 0.48±0.09 | 0.22±0.03 | 0.27±0.02 |
| F1-score | 0.28±0.01 | 0.41±0.01 | 0.49±0.11 | 0.25±0.03 | 0.28±0.03 |

*Table 5 Test Metrics for ResNet152*

Performance of test (Table 5) confirms the solidity of this backbone. F1-scores remain relatively high for necrosis (0.49 ± 0.11) and infiltration (0.41 ± 0.01), but even cdiff, a problematic descriptor, is at 0.28 ± 0.01 and scores better than the other extractors within this class. Standard deviations are also generally lower overall, indicating ResNet152 has more consistent performance across validation folds.

### 8.1.3 MobileNetV2

The retrieval loss curve in Figure 13 indicates a good training process. The loss quickly drops from ~2.9 to 2.75 initially as the model begins learning general retrieval habits. Then, the decline is smooth with small oscillations as is typical when stochastic optimization is employed but which do not hinder convergence. From epoch 10, the

loss plateaus at 2.6, and then decreases again to around 2.45 at epoch 23.

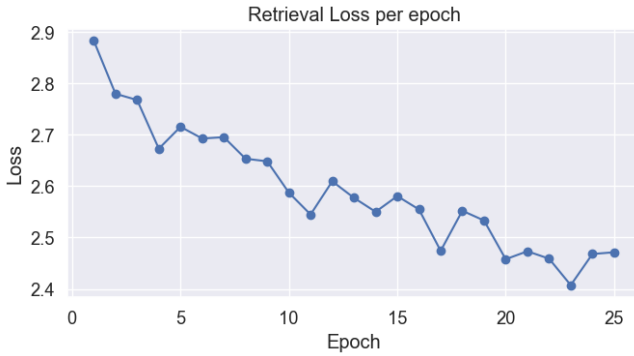


*Figure 12 MobileNetV2 Loss Curve*

MobileNetV2 testing results are presented in Table 6 (train) and 7 (test). The thin architecture tends to have lower and more inconsistent performance compared to the other backbones, especially when generalizing to unseen data.

| Test | cdiff | infiltration | necrosis | nodule_density | nodule_shape |
|---|---|---|---|---|---|
| recall | 0.25±0.15 | 0.31±0.15 | 0.48±0.38 | 0.27±0.27 | 0.28±0.18 |
| precision | 0.22±0.13 | 0.30±0.14 | 0.47±0.38 | 0.27±0.24 | 0.30±0.14 |
| F1-score | 0.22±0.12 | 0.29±0.13 | 0.47±0.38 | 0.25±0.24 | 0.26±0.12 |

*Table 7 Test Metrics for MobileNetV2*

On training (Table 6), the model performs well for necrosis with an F1-score of $0.90 \pm 0.01$ and moderately for infiltration ($0.52 \pm 0.06$). However, descriptors like cdiff and nodule shape have much lower scores (about 0.43 and 0.42 respectively), suggesting that the model cannot describe more abstract or less visually differentiated classes.

In the test period (Table 7), performance is substantially reduced for cdiff (F1-score: $0.22 \pm 0.12$) and for nodule shape ($0.26 \pm 0.12$), with considerable variance across folds. Necrosis remains the most predictive descriptor with consistency ($0.47 \pm 0.38$), although with instability indicated by standard deviation.

In summary, MobileNetV2 depicts clear deficiencies in learning generalizable representations for this task. While its reduced computational cost is enticing, the reduction in predictive reliability, especially in underrepresented domains, suggests that more powerful extractors may be needed for medical imaging tasks requiring fine-grained clinical classification.

## 8.2 Top-k Performance

This experiment contrasts the way the retrieval performance of the model varies with different top-k values, using ResNet152 as the visual feature extractor. For each clinical descriptor, a correct prediction is assigned to the model if the true value falls in any of the top-k returned reports. This is measured in terms of accuracy per descriptor with **success counted by whether the correct value is present in one of the top-k results**.

As seen in Figure 14, all descriptors exhibit a consistent increase in results when k is increased. This trend confirms that the model tends to retain crucial information within its greater collection of most similar outputs, albeit not necessarily always having the correct value ranked first.

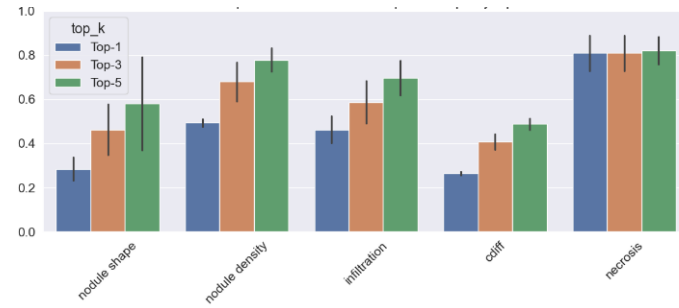


*Figure 13 Top-k Accuracy per Descriptor*

The density of nodule gets much better from top-1 at 49% to top-5 at 77%. This almost 30-point gap shows that even though the correct value may not be in the model's top but sometimes it does appear there among its top guesses. Similarly, nodule shape—which as such an ever-so-subjective, variable descriptor is less likely to be represented accurately—starts lower at top-1 29% but rises to top-5 57%. This indicates that the model is capturing correct visual information but with some ambiguity and decreased confidence in ranking.

On the other hand, necrosis performs extremely well even in top-1 scenarios with 81% recall per descriptor and does not improve much with more general top-k, which means that it is a more salient and unambiguous feature. Meanwhile, terms like cdiff and infiltration, beginning at low top-1 results (30%–47%), obviously increase as k increases. This shows that although the model is capturing significant content, it is not necessarily retrieving the correct value at the top, demonstrating the value of top-k evaluation to reveal hidden predictive capability.

Overall, these results emphasize the use of top-k evaluation in image-to-report retrieval tasks and provides a better understanding of the behavior.

## 8.3 Impact of contrastive loss

The Triplet Loss experiment training curve (Figure 15) shows a huge fall in the initial epochs, characteristic of good early-stage learning. The loss settles instantly at 0.200, showing that the model converges well. Compared to the CrossEntropyLoss-trained ResNet152 model, the Triplet Loss experiment also shows a more conclusive drop-off, though not quite as smooth, which may show a more linear optimization trajectory in embedding space.

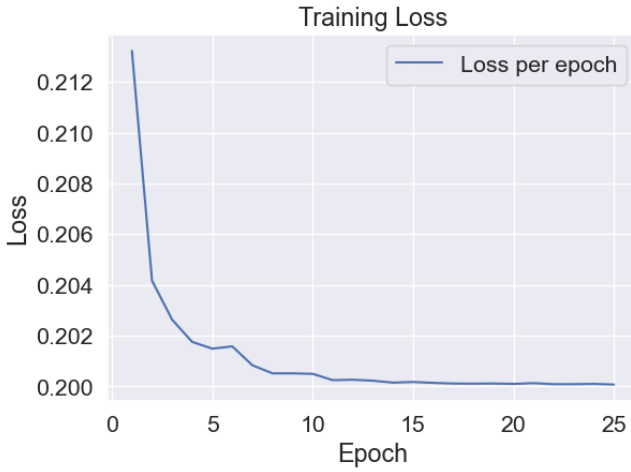


*Figure 14 Triplet Margin Loss Curve*

Figure 16 illustrates the fact that the accuracy per descriptor is improved at larger top-k values, particularly for cdiff and nodule shape, with best performance at top-5. Necrosis remains stable across top-k values. Compared to ResNet152, the model based on Triplet Loss generally has higher accuracy across all the top-k levels, indicative of improved fine-tuning, though ResNet152 is steadier, especially at top-1.

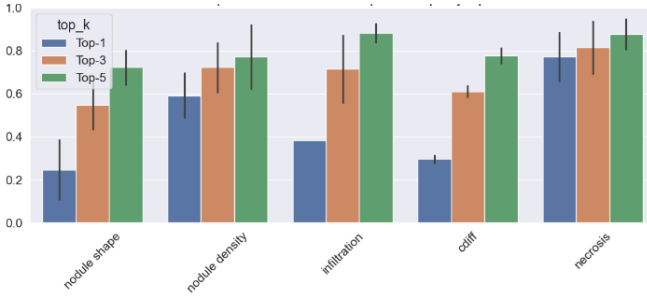


*Figure 15 Per Descriptor Accuracy*

### 8.4 Nodule respresentation space comparision

Here, in the forth experiment, we add an attention mechanism to aggregate visual features from throughout different CT slices. Instead of averaging, the model learns importance weights for all of the subset of 15 selected slices and places greater emphasis on those most important for matching against clinical text. This attention module is trained in conjunction with the rest of the model on the main loss function so that it can learn to focus on the most informative regions of the scan.

Although the loss observed in Figure 17 tends to go down overall through training epochs, the curve only shows moderate improvement. The decline is relatively slow, and the ultimate plateau of the loss remains comparatively high, meaning that attention-based model is unable to achieve robust correspondence between text and image representation.

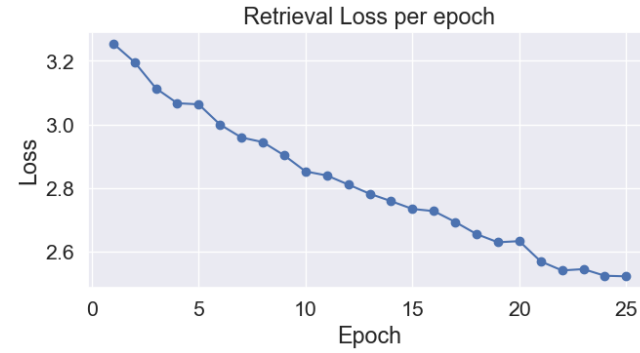


*Figure 16 Loss Curve with Attention Mechanism in visual features aggregation*

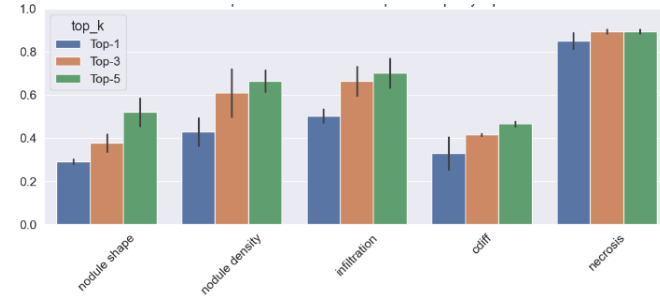


*Figure 17 Per Descriptors Accuracy*

Per descriptor accuracy values obtained in Figure 18 through the attention-based aggregation are mostly lower than those obtained from previous experiments. Although some descriptors like necrosis do not suffer significantly with a change in top-k, others like nodule shape or cdiff remain less than 50% even at the top-5 scenario. This may indicate that the attention mechanism, although adaptive, is not presently a major advantage over simpler aggregation techniques in this task.

## 9 Conclusion

This project investigated if the CLIPMedical model could project structured clinical descriptors onto CT-derived visual features in a supervised contrastive framework. The main findings from each experiment are as follows:

In ***Visual representation space comparision experiment***, the test illustrated that increased backbones drastically improve model performance. ResNet152 outperformed ResNet18 and MobileNetV2 with greater and more stable F1-scores, particularly for descriptors like necrosis and infiltration. This signifies the importance of rich visual representations to accurate image-text alignment.

In the ***Top-k performance analysis***, per descriptor ac-

curacy also increased consistently with a shift from top-1 to top-5 retrievals. Features such as nodule shape and density, otherwise visually ambiguous, benefited from this versatility. This shows that top-k evaluation provides an improved estimate of the interpretability potential of the model in actual clinical practice.

In the ***Impact of contrastive loss*** experiment, training improved retrieval per descriptor accuracy, especially for difficult descriptors like cdiff. Convergence was not as smooth as with cross-entropy, indicating that while triplet loss increases fine-grained structure in the embedding space, it requires careful tuning and sampling strategies.

In the ***nodule space representation comparision***, slice selection attention did not beat the baseline mean pooling. That is, the attention module in this setup was not able to always assign strong preference to key diagnostically relevant slices, likely due to sparsity or corruption of the feature set.

A patient breakdown established that while there were some instances where cases generated ideal descriptor predictions, most fell within the range of 40% to 80% accuracy. This demonstrates partial semantic matching but also establishes how, currently, the model is unable to generate full and clinically relevant summaries on an ongoing basis.

Overall, this research is an early exploration of contrastive multimodal learning for explainable medical retrieval. Although *CLIPMedical* is promising, current performance indicates severe limitations—mainly class imbalance and noisy labels—that place it far beyond clinical use. There is much to be done in terms of data quality and model robustness before real-world deployment.

## ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my tutor, Debora, for her guidance, advice, and continuous support throughout the development of this project. I would also like to thank my family and my couple for their support, patience, and motivation throughout these past months. I'm especially thankful to all the friends who have shared this journey with me, and to all the professors whose dedication and effort helped shape us into the engineers we are becoming.

## BIBLIOGRAFIA

[1] D.S. Ardila, A.P. Kiraly, S. Bharadwaj, et al., "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, pp. 954–961, 2019.

[2] A.A.A. Setio, F. Ciompi, G. Litjens, et al., "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1160–1169, May 2016.

[3] A. Holzinger, B. Malle, and A. Saranti, "Towards Multi-modal Causability with Graph Neural Networks enabling Information Fusion for Explainable AI," *Information Fusion*, vol. 71, pp. 28–37, Feb. 2021.

[4] G. Tjoa and W. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[5] A. Radford, J. Kim, et al., "Learning Transferable Visual Models from Natural Language Supervision," OpenAI, https://openai.com/research/clip, 2021.

[6] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[7] Y. Zhang, Y. Yang, L. Zhang, J. Gao, and H. Xu, "When Radiology Report Generation Meets Knowledge Graph," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 260–269, 2023.

[8] Y. Jing et al., "Automatic Radiology Report Generation: A Survey," *arXiv preprint*, https://arxiv.org/abs/2006.13657, 2020.

[9] X. Chen et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *Proc. ICML*, pp. 2048–2057, 2015.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," *Proc. CVPR*, pp. 3156–3164, 2015.

[11] R. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *Proc. ICLR*, 2015.

[12] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proc. ICLR*, 2021.

[13] Y. Zhang et al., "Contrastive Vision–Language Pretraining for Radiology: A New Benchmark and Dataset," *Nature Machine Intelligence*, vol. 5, no. 5, pp. 398–407, May 2023. (BiomedCLIP)

[14] T. L. Huang et al., "GLORIA: A Multimodal Global-Local Representation Learning for Medical Images and Reports," *Proc. CVPR*, 2021.

[15] K. Sohn, "Improved Deep Metric Learning with Multi-class N-pair Loss Objective," *Proc. NeurIPS*, 2016.

[16] T. Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations," *Proc. ICML*, pp. 1597–1607, 2020.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv preprint*, https://arxiv.org/abs/1807.03748, 2018.

[18] D.J. Anderson, *Kanban: Successful Evolutionary Change for Your Technology Business*, Blue Hole Press, 2010.

[19] Atlassian, "What is Trello?", https://trello.com/guide, 2023.

[20] H. Haixiang et al., "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 1–16, Jan. 2017.

[21] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley, 2007.

[22] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification*, vol. 2, no. 4, pp. 193–218, Apr. 1985.

[23] E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings," *Proc. Clinical NLP Workshop (NAACL)*, pp. 72–78, 2019.

[24] H. Noh et al., "Image Retrieval with Deep Local Features," *Proc. CVPR*, 2017.

[25] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *J. Mach. Learn. Tech.*, vol. 2, no. 1, pp. 37–63, 2011.

[26] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. IJCAI*, pp. 1137–1145, 1995.

# APÈNDIX

## A1. FREQUENCY TABLE REFERED IN SECTION 5

| Descriptor | | Count |
|---|---|---|
| cdiff | | |
| | missing | 76 |
| | moderatelly differentiated | 61 |
| | well differentiated | 27 |
| | poorly differentiated | 18 |
| nodule_density | | |
| | Solid | 115 |
| | Part-Solid | 53 |
| | Groundglass | 12 |
| | missing | 2 |
| nodule_shape | | |
| | Spiculated | 69 |
| | Irregular | 55 |
| | Round/Oval | 35 |
| | Lobulated | 23 |
| vinfiltration | | |
| | No | 93 |
| | missing | 79 |
| | Yes | 10 |
| necrosis | | |
| | 0.0 | 168 |
| | 20.0 | 14 |

*Table 2 Frequence Table of each label*

## A2. PROJECT MANAGEMENT WITH TRELLO

To sequence and schedule the workflow of this project, I utilized the Kanban method using the Trello tool. The tool was straightforward to use in sequencing activities into three general categories: "To Do", "In Progress", and "Done", as shown in Figure 20. This was more convenient for me to apply in task prioritization, tracking progress efficiently, and adapting to any unexpected variations in timing or scope..

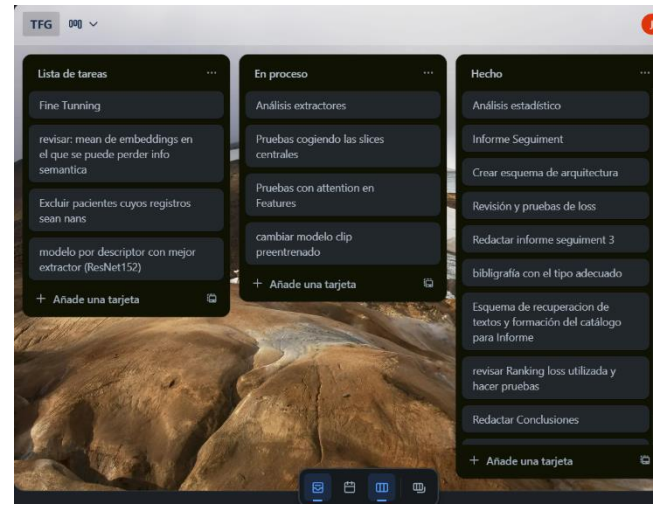


*Figure 18 Screenshot of the Trello board used for task management*

## A3. EXTRA EXPERIMENT: PER-PATIENT DESCRIPTOR ACCURACY ANALYSIS

Following the comparative evaluation of the different image feature extractors, ResNet152 was selected for further analysis as it showed slightly better overall performance. To gain a more realistic understanding of the system's clinical applicability, a per-patient analysis was performed based on the number of correctly predicted descriptors per case.

Here, it is calculated the proportion of correctly predicted descriptors per patient. This enables it to identify how often the model produces entirely or mainly correct clinical summaries, which is especially significant in real-world contexts.

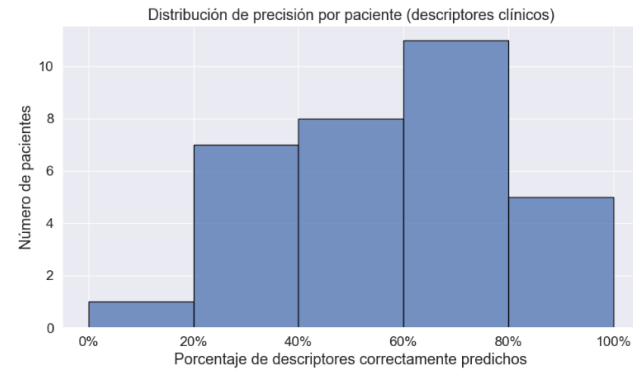


*Figure 19 % Accuracy Descriptors per Patient*

The results in Figure 19 show that there were some patients who obtained all predicted descriptors correct (100%), but most fell between 40% and 80% correct. This suggests that the model is capturing important information but is often not generating complete and coherent outputs. It doesn't seem like there are groups of descriptors which fail together consistently or whether there are specific clinical profiles that have consistently poorer predictions.

## A4. TRAIN METRICS TABLES

Here are presented the refered train metrics tables from the Experiment 1.

| Train | cdiff | infiltra-tion | necrosis | nod-ule_densi ty | nod-ule_sha pe |
|---|---|---|---|---|---|
| recall | 0.55±0.0 | 0.47±0.06 | 0.60±0.0 | 0.39±0.043 | 0.44 ± 0.0 |
| preci-sion | 0.60±0.1 | 0.46±0.06 | 0.59±0.14 | 0.56±0.057 | 0.433±0.0 |
| F1-score | 0.55±0.0 | 0.46±0.05 | 0.59±0.00 | 0.38±0.05 | 0.435±0.0 |

*Table 2 Train Metrics for ResNet18*

| Train | cdiff | infiltration | necrosis | nodule_density | nodule_shape |
|---|---|---|---|---|---|
| recall | 0.56±0.01 | 0.47±0.03 | 0.81± 0.01 | 0.51±0.10 | 0.49±0.09 |
| precision | 0.42±0.02 | 0.46±0.02 | 0.58±0.00 | 0.38±0.08 | 0.33±0.13 |
| F1-score | 0.48±0.00 | 0.56±0.02 | 0.68±0.02 | 0.46±0.1 | 0.42±0.12 |

*Table 4 Train Metrics for ResNet152*

| Train | cdiff | infiltration | necrosis | nod-ule_density | nod-ule_shape |
|---|---|---|---|---|---|
| recall | 0.43±0.10 | 0.51±0.02 | 0.91± 0.01 | 0.45±0.22 | 0.44±0.13 |
| precision | 0.52±0.19 | 0.68±0.12 | 0.91±0.02 | 0.56±0.13 | 0.46±0.14 |
| F1-score | 0.43±0.12 | 0.52±0.06 | 0.90±0.01 | 0.40±0.16 | 0.41±0.13 |

*Table 6 Train Metrics for MobileNetV2*

## A5. FUTURE IMPROVEMENTS AND NEXT STEPS

Future improvement should focus on dataset balancing, noisy or missing label curation, and descriptor annotation enrichment to accommodate more clinical variability. Pretraining and multi-task learning methods can also help the model identify more text-imaging correlations. All these steps are essential to move beyond exploratory results towards clinically valid, more generalizable results.

Also, it would be interesting to explore the approach of creating individual embeddings for each descriptor instead of combining all the descriptors into a single embedding as in this project. Processing descriptors individually could provide a more detailed understanding of the connection of each feature with the medical images so that the model can identify specific patterns for each diagnostic criterion more effectively. This would be able to improve the system overall accuracy and interpretability, offering a more flexible means for future improvements.

## A5. RESNET152 CONFUSION MATRIX

Here it is the confusion matrix refered and explained in the ResNet152 paragraph within the Experiment 1.

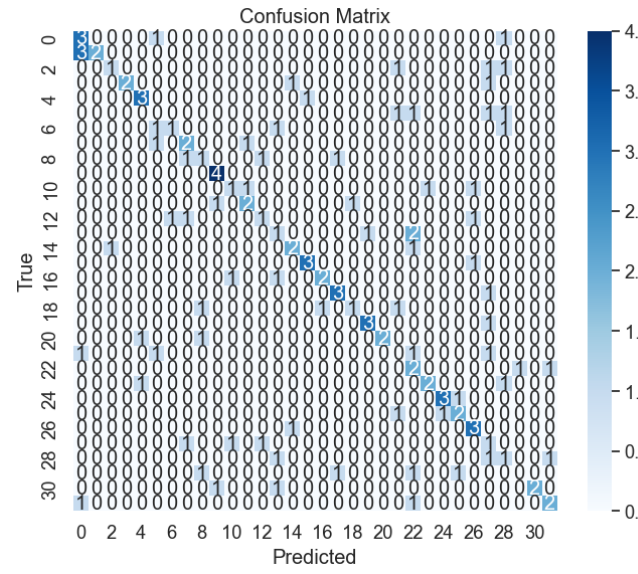


*Figure 20 Text-Image Embeddings Confusion Matrix*

## A5. DEVELOPMENT SCHEMAS

The following are the diagrams explained and referenced in section 6, Development, expanded for better detail.

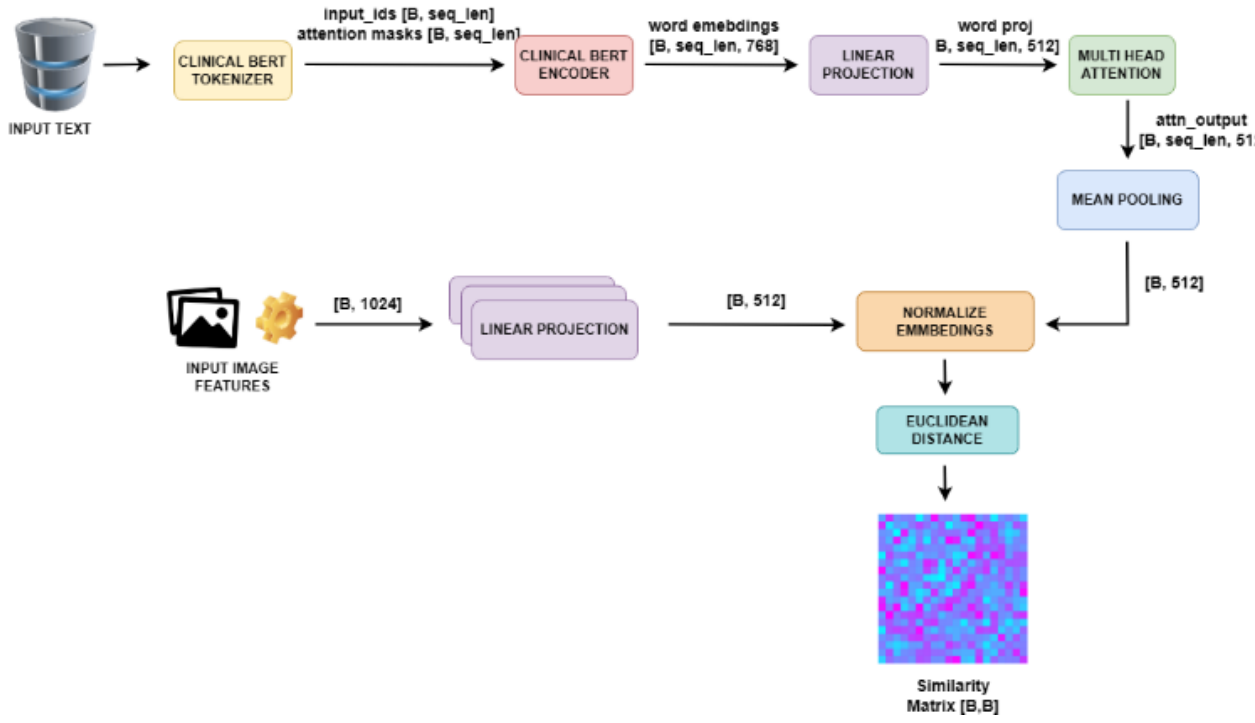


*Figure 7 Architecture Schema*

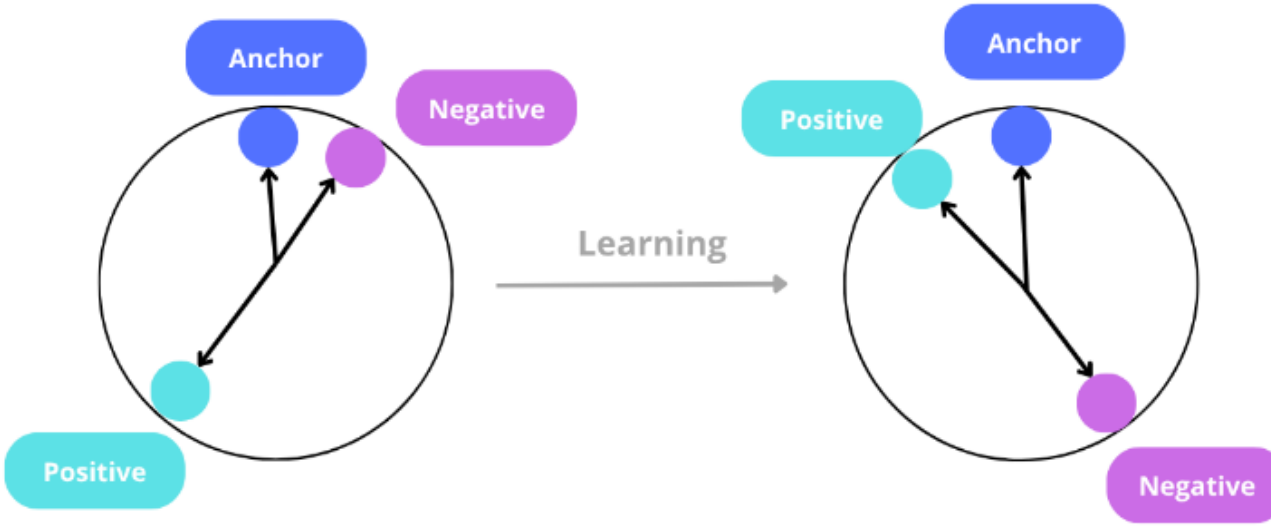


*Figure 8 Triplet Margin Loss Schema*

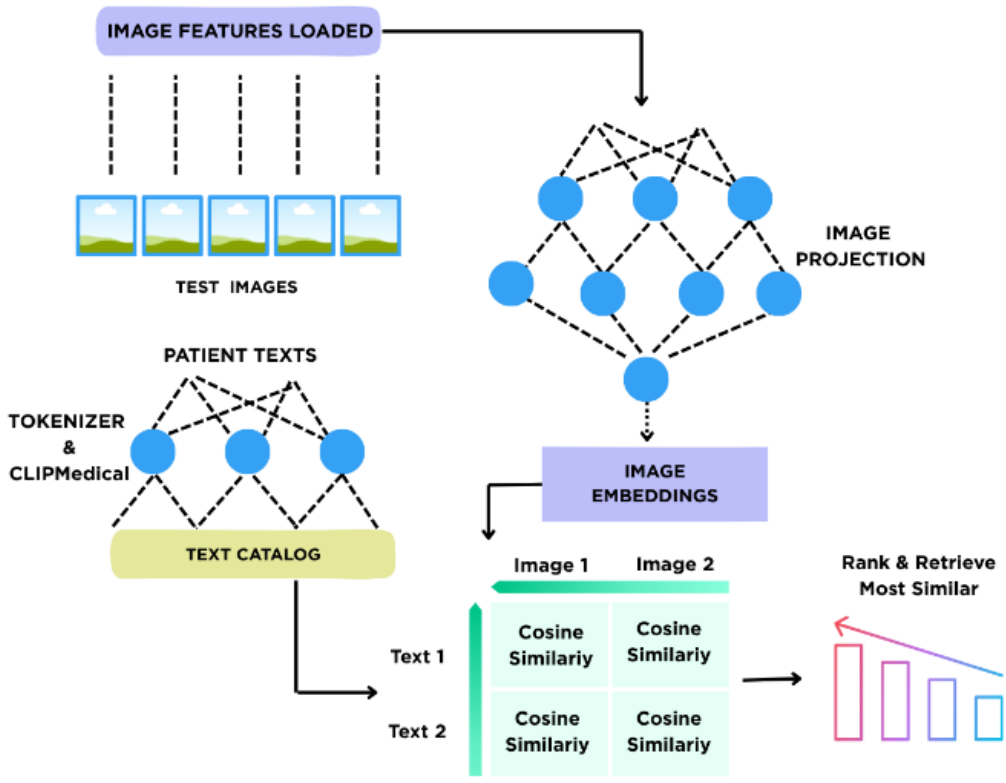


*Figure 9 Inference and evaluation schema*