

# MMG-CLIP: Automated Mammography Reporting through Image-to-Text Translation

Abdelrahman Habib<sup>a</sup>, Santiago Pires<sup>b</sup>, Jaap Kroes<sup>b</sup>

<sup>a</sup>Universitat de Girona, Spain; University of Bourgogne, France; Università degli studi di Cassino e del Lazio Meridionale, Italy

<sup>b</sup>ScreenPoint Medical, Nijmegen, Netherlands

---

## Abstract

Recently medical image-text datasets have become increasingly important in the development of deep learning applications, including automated radiology report generation models. Generating clinically valid radiology reports comes along with challenges, such as bridging the gap between interpreting medical images and accurately conveying the findings into radiology text reports. In this work, we tackle the task of automated mammography report generation following Breast Imaging Reporting & Data System (BI-RADS) guidelines. We utilize an image-label and exam-reports datasets, along with text prompting techniques, to generate a well-structured text report that supports training. Our proposed framework allows the usage of up to four image views within the exam, leveraging different information that can be captured from all exam views related to the radiology report. Our model demonstrated high performance in supervised and zero-shot classification settings when evaluated on multiple downstream tasks, enabling report generation as a series of zero-shot classification tasks.

**Keywords:** Mammography 2D X-ray, BI-RADS Report Generation, Contrastive Learning, Natural Language Processing

---

## 1. Introduction

Medical images from different modalities such as Mammography X-ray, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) are widely used to evaluate, monitor, and diagnose several medical conditions in clinical practice. Mammography X-ray is a universally accepted method for breast cancer detection as it is relatively inexpensive, repeatable, and widely available (Fishman and Rehani, 2021). Several applications demonstrated the effectiveness of deep-learning based models on solving tasks related to breast cancer detection in mammography, such in discrimination of microcalcifications (Wang et al., 2016), microcalcifications detection (Pesapane et al., 2023), breast cancer risk discrimination (Yala et al., 2019), and breast cancer image segmentation (Salama and Aly, 2021), and many others (Kallenberg et al., 2016; Mohamed et al., 2018; Ribli et al., 2018).

Although deep-learning models, such as convolutional neural networks (CNNs) by He et al. (2016); Krizhevsky et al. (2012); Simonyan and Zisserman

(2014) have been widely applied for various artificial intelligence (AI) tasks in recent years (Han et al., 2021), and has been actively used for the purpose of medical image analysis (Anwar et al., 2018), the small size of annotated and publicly available medical datasets remains a major bottleneck in this area for developing computer-aided detection/diagnosis (CAD) tools. Unlike publicly available computer vision dataset that are available in large-scale, such as ImageNet (Deng et al., 2009) or OpenImages (Kuznetsova et al., 2020), publicly available medical datasets are much smaller in magnitude (Xie et al., 2021). This introduces challenges in training deep-learning models for medical purposes as the availability of high-quality clinical annotations is time-consuming and costly (You et al., 2023), and obtaining labels for medical images is very resource-intensive as it relies on domain experts (Karimi et al., 2020). Therefore, building effective medical imaging models is limited by the lack of large-scale annotated medical dataset.

Recently, Contrastive Language-Image Pre-training

(CLIP) as in the work of Radford et al. (2021), has achieved considerable success in computer vision and natural language processing domains, by allowing joint-training of image and text representation on large-scale image-text pairs (Wang et al., 2022), enabling zero-shot transfer of the model to downstream tasks. As shown by Radford et al. (2021), zero-shot CLIP models are much more robust than equivalently accuracy supervised ImageNet models. In another work, ALIGN by Jia et al. (2021) similarly to CLIP trains dual-encoder architecture to learn the alignment of visual and language representations of image and text pairs using contrastive loss by leveraging noisy dataset of over one billion image alt-text pairs. Both ALIGN and CLIP shows great robustness on classification tasks with different image distributions (Jia et al., 2021).

Considering CLIP, adopting such large vision-text pre-training models to the medical domain is a non-trivial task due to CLIP’s data-hungry nature that was trained on 400 million (image, text) pairs collected from the internet (Wang et al., 2022). In that context, the natural solution of limited annotated medical dataset is to leverage the corresponding medical reports that contain detailed description of the medical condition observed by radiologists (Huang et al., 2021).

## 2. State of the art

### 2.1. Contrastive learning approaches

Several recent works to utilize both medical images and text in the domain of chest X-ray (Huang et al., 2021; Li et al., 2021; Wang et al., 2022; You et al., 2023), using CLIP-based architecture. GLoRIA framework by Huang et al. (2021) uses an attention mechanism by contrasting image sub-regions and words in the paired report by learning attention weights that emphasize significant image sub-regions for a particular word to create context-aware local image representation. MedCLIP by Wang et al. (2022) on the other hand used unpaired images, text, and labels to enhance medical multi-modal learning. However this makes it less capable of retrieving the exact report for a given image due to the effect of decoupling image-text pairs, and as their approach relies on the performance of their rule-based labeler, it is not scalable to other diseases that the labeler can’t address (You et al., 2023).

DeCLIP by Li et al. (2021) introduced a novel paradigm for data efficient CLIP that tackles the limitation of training data availability similar to the amount that CLIP was trained on through (1) self-supervision within each modality, (2) multi-view supervision across modalities, and (3) nearest-neighbor supervision from other similar pairs. CXR-CLIP by You et al. (2023) utilizes both image-text pairs not only from image-text dataset, but also from image-label dataset, thus tackles the lack of image-text data in the chest X-ray domain

by expanding image-label pair via general prompting. In their work, they also used Multi-View Supervision (MVS) as inspired by Li et al. (2021), utilizing multiple images and texts in a chest X-ray study, such as two distinct images and texts pairs each using an augmentation approach.

### 2.2. Convolutional neural network approaches

Other approaches have utilized convolutional neural networks in generating medical image descriptions or reports (Jing et al., 2017; Kisilev et al., 2016; Wang et al., 2018). In the work of Kisilev et al. (2016), they trained a CNN-based architecture to generate and rank rectangular region of interests of breast mammography and ultrasound modalities, where highest score candidates are fed to the subsequent network layers, in which they are trained to generate semantic description of the remaining ROI’s. Their network is based on Faster R-CNN architecture (Ren et al., 2015), and was trained on mini-batches of positive and negative ROI candidates, and requires rectangular ground truth bounding boxes. Their main goal was to test the description stage of images using some descriptors such as mass shapes and margins.

Other approaches as Jing et al. (2017) utilized a hierarchical Long-Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997), apart of a multi-task learning framework to generate long report paragraph in chest X-ray domain. TieNet by Wang et al. (2018) is a multi-purpose text-image embedding network that utilizes report data together with paired images to produce meaningful attention-based image and text representations in the chest X-ray domain. Their approach also uses the paired text-image representations from training as a *priori* knowledge injected, to improve classification and generate text reports. They introduced an attention encoded text embedding mechanism that provides more meaningful text embedding, tackling the challenge that comes along with long reports of multiple information.

### 2.3. Limitations of current methods

Despite such novel contributions made in the medical imaging chest X-ray domain using medical image-text datasets, several challenges still exist in the mammography X-ray domain, and specifically for BI-RADS report generation, which are summarized as follows:

- **Complications of mammography text reports.** Most of the present work utilizes chest X-ray image-text datasets, where the paired reports could be summarized under “impressions” and “findings”, making it easy to extract text information for training. Mammography text reports on the other hand could contain additional information to the X-ray radiologist report that can be used as gold standard confirmation, such as ultrasound, MRI, or

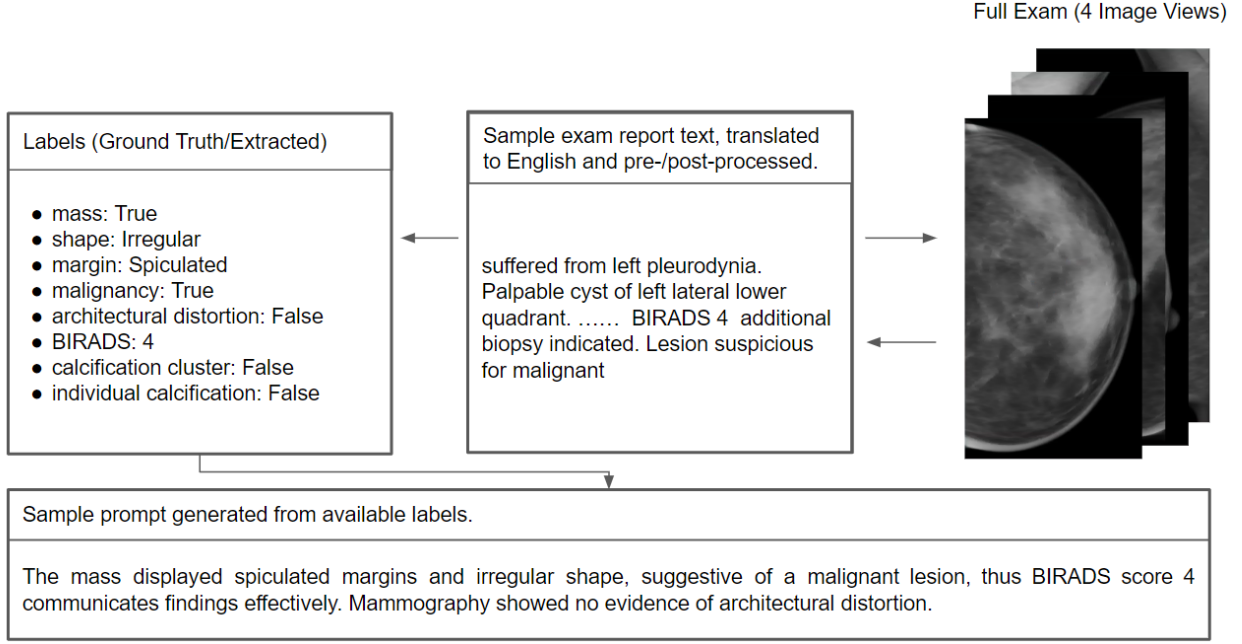


Figure 1: Example of mammography visual exam paired to different structure of text information, such as text report, extracted labels, or a prompt generated sentences using a text template based on the available labels for the exam.

pathology reports. Those additional information introduces challenges in identifying the best section for training a network. For instance, mammography X-ray radiology report could indicate suspicious morphology for a study, however, malignancy is confirmed by a biopsy and from an MRI exam. Such information can be mentioned in the same report of a single exam study, making it more challenging for the network to understand the meaning of different sections available in the patient reports.

- **Pathology variability in different views.** Unlike the work that is presented by CXR-CLIP (You et al., 2023), which utilizes up to two views with augmentation, mammography X-rays could contain up to four views (two for each breast - mediolateral oblique (MLO) and cranial caudal (CC)). With that, it could be possible to have a specific pathology in one breast and not in the other, increasing the necessity of having a network that is capable to process all four exam views and pair them to the text dataset.
- **Limited available data.** Most image-text datasets which are publicly accessible are available for different domains as chest X-ray (Bustos et al., 2020; Johnson et al., 2019), unlike mammography X-ray. And as the nature of its radiology reports, it is even more difficult to find paired images and full text reports, leaving a vast majority of image-label datasets unused to tackle the report generation task.

#### 2.4. Contributions of this work

The main contribution of this work is summarized as follows:

1. To our knowledge, this is the first work to utilize CLIP approach in mammography X-ray domain for mammography report generation. We tackle the lack of data by utilizing image-label and exam-reports paired datasets, as well as generating text prompts based on available labels to support the training. Our method, namely *MMG-CLIP*, does not depend on a ruler-based labeler, and doesn't require bounding boxes or small-patched images for training, and can be adapted to any image-label or exam-reports dataset.
2. We implemented a training approach that utilizes four views per exam, pairing them to the same text description, whether a label, a generated text prompt, or a report used during training or evaluation.
3. Performance of our model is validated on multiple downstream classification tasks, using zero-shot and supervised classification settings, as well as measuring the performance with respect to data-efficiency.
4. We introduced the report generation pipeline as a series of zero-shot classification tasks following BI-RADS guidelines, to obtain a clinical meaningful draft report for the patient exam.

### 3. Material and methods

The aim of this work is to learn a multi-modal embedding space from features that are extracted from an image and text encoders, and projected to a similar embedding dimension, to maximize the cosine similarity of both image and text embedding of real pairs in each batch, and minimize the cosine similarity of the incorrect embedding pairings, similarly to CLIP (Radford et al., 2021). Our approach aims to learn the image level or exam level characteristics of the 2D mammography X-ray images, up to four image views per exam. Those characteristics are also sampled from both image-label and exam-report datasets, in addition to the prompt generation approach to support training. In the following subsections, we further explain our work.

#### 3.1. Data Sampling

To train the model, each batch consists of both visual and textual information. Similarly to the work presented by You et al. (2023), we utilize a set of images, however, each exam could contain up to four image views. Thus, each batch sample consists from one to four  $X_{\text{img}}$  images depending on their availability for each exam, and  $T_{\text{txt}}$  text. To simplify the following explanation, we denote quantities related to the full exam as  $X_{\text{exam}}$  as in equation 1.

$$X_{\text{exam}} = \{X_{\text{img}}\}_{\text{img}=1}^4 \quad (1)$$

In the case of image-label dataset, the sampled text  $T_{\text{txt}}$  could be the an exact single label, for instance "benign" or "malignant" labels. Also, we use such labels, with any other labels found for the image to generate prompts that supports the model training. Those prompts we used contains more than one class label information, unlike the work of You et al. (2023) that only consists of one class-specific information. We also considered cases where the image-label pairs are missing labels information, making the prompts close to real clinical reports and taking into account not only the class information but their appearance.

For the exam-report dataset, the sampled text  $T_{\text{txt}}$  consists of the processed report information, using certain selected reports sections found in the report text. In addition to that, as we had labels for the exams, we also experimented the training performance with generated sentences based on labelled data, known as prompts, and with both reports and prompts combined. We demonstrate a sample of image-prompt pairs from the training set in Appendix A, where we used our prompts as text input for training. We also took into account that those prompts are applicable with the BI-RADS guidelines and information that can be extracted from it. Figure 1 demonstrates different types of mammography datasets. Further details on the dataset and prompting mechanism is elaborated in subsection 4.1.

#### 3.2. Model Architecture

Motivated by CLIP by Radford et al. (2021), we proposed slight modification to how the embedding are extracted from multiple exam views to allow processing more than one mammography X-ray image at one time, as well as text feature extraction, both are described in subsections 3.2.1 and 3.2.2. In subsection 3.2.3, we describe the projection approach, that is necessary to align the embedding to the same dimension. Finally, subsection 3.3 describes the loss term that trains the model. All of this is summarized in Figure 2.

##### 3.2.1. Image Encoder

The image encoder was used extract features from each exam input image, where the encoder can be referred to as in the following equation 2.

$$x = E_{\text{img}}(X_{\text{img}}) \quad (2)$$

where  $x \in \mathbb{R}^{1 \times D_{\text{img}}}$  represent the feature vectors for a single image view, and  $E_{\text{img}}$  represents the image encoder. This is repeated for  $N$  number of exam views, denoted as  $x_{\text{exam}}$  where  $x_{\text{exam}} \in \mathbb{R}^{N \times D_{\text{img}}}$ . The value  $D_{\text{img}}$  is the dimension of each vector. To obtain an overall visual representation of the exam, we average the values of all feature vectors of all exam views along the 0-th dimension, denoted at  $x_f$ , which is computed as following.

$$x_f = \frac{1}{N} \sum_{i=1}^N x_{\text{exam}}(i) \quad (3)$$

where  $x_{\text{exam}}(i)$  represents the  $i$ -th column of matrix  $x_{\text{exam}}$ . The resulting  $x_f$  has shape  $(1, D_{\text{img}})$  representing the final image embedding vector. In the case that the network is trained at the image level where the input consists of a single image paired with the text, the averaging process is not performed and equation 2 is denoted as  $x_f$ .

The image encoder we used is a ConvNeXt Tiny model (Liu et al., 2022), pre-trained on an internal multi-vendor dataset from Fujifilm, GE HealthCare, HOLOGIC, Lorad, Philips and Siemens Healthineers, on large-scale dataset (>100K exams) for malignancy classification. In addition to that, we used ResNet-50 model from He et al. (2016) with ImageNet weights pre-trained on ImageNet tasks (Deng et al., 2009) in our ablation study to assess the performance when using a domain-specific pre-trained model to other pre-trained models.

##### 3.2.2. Text Encoder

The text encoder was used to extract features from the input text. It can be described as the following equation 4.

$$t_f = E_{\text{txt}}(T_{\text{txt}}) \quad (4)$$

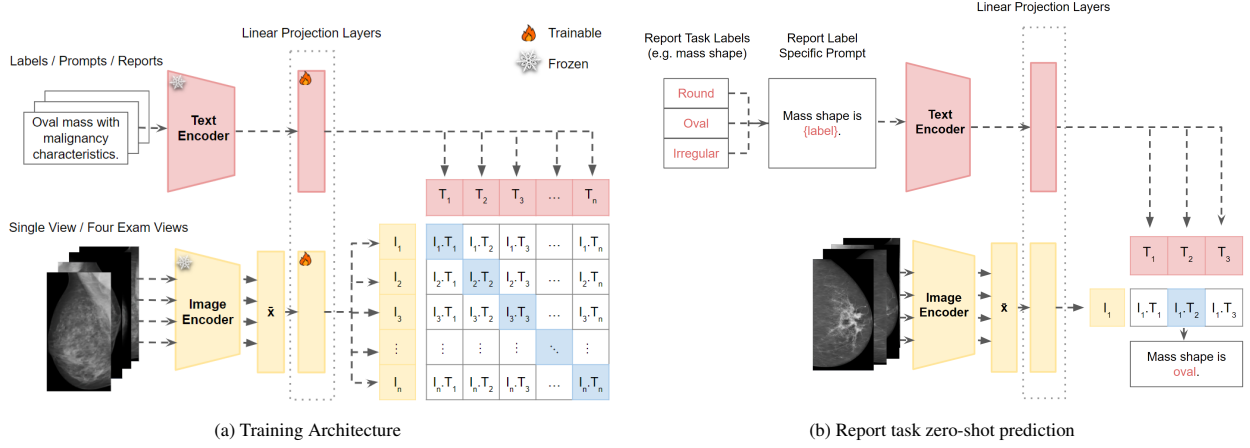


Figure 2: Summary of our approach motivated by CLIP (Radford et al., 2021). *MMG-CLIP* extracts features from both text and image view/exam views, averages the image embedding, and projects them to predict the correct pairings of each batch. At inference, the network outputs unnormalized probability distribution for the input texts representing their probability to be paired to the input image. We aim to utilize this approach in report generation where a draft report is generated as a sequence of zero-shot classification tasks based on BI-RADS guidelines.

where  $t_f \in \mathbb{R}^{1 \times D_{\text{txt}}}$  represents the text embeddings and  $E_{\text{txt}}$  represents the text encoder. We used BioClinicalBERT model by Alsentzer et al. (2019), which is a Bidirectional Encoder Representations from Transformers (BERT) based model as our text encoder, that was pre-trained using clinical dataset MIMIC-III (Johnson et al., 2016), similar to (Huang et al., 2021; Wang et al., 2022; You et al., 2023).

We also used BiomedBERT previously named as PubMedBERT (Gu et al., 2021), and BioGPT by Luo et al. (2022) to compared the performance when using BioClinicalBERT in our ablation study as in section 5. BiomedBERT is also a variant of BERT models (Devlin et al., 2018), that was pre-trained from scratch on data collection from PubMed<sup>1</sup> that consists of 14 million abstracts and 3.2 billion words. This model was pre-trained on biomedical domain-specific data compared to BERT that is trained on Wikipedia<sup>2</sup> and BookCorpus (Zhu et al., 2015) as cited in (Gu et al., 2021). BioGPT is a variant of GPT large language models (LLMs), that is a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature for biomedical text generation and text mining (Luo et al., 2022). It was pre-trained on 15M PubMed abstracts from scratch on GPT-2 (Radford et al., 2019) model configuration as a backbone, thus resulting into a model with 0.355 billion parameters in total as cited in Luo et al. (2022). In our experiments, we used all of the pre-trained text encoders from HuggingFace<sup>3</sup>.

### 3.2.3. Embedding Projection

To align both the image embedding  $x_i$  and text embedding  $t_f$  in the same multimodal feature space, we trained linear layers as projection heads.

$$v = \frac{f_x(x_i)}{\|f_x(x_i)\|} \quad (5)$$

$$u = \frac{f_t(t_f)}{\|f_t(t_f)\|} \quad (6)$$

where  $f_x$  is the projection head for the image embedding,  $f_t$  is the projection head for the text embedding,  $v$  and  $u$  are the normalized projected embedding,  $V = \{v\}_{i=1}^n$ ,  $U = \{u\}_{i=1}^n$ , and  $n$  is the batch size.

### 3.3. Loss Function

For the loss, CLIP utilizes InfoNCE loss by Oord et al. (2018) as cited in Li et al. (2021), which is a symmetrical loss for image and text encoder. It iteratively trains both image and text encoders to maximize the cosine similarity of the image and text embedding of the  $N$  real pairs in the batch, while minimizing the cosine similarity of the image and text embedding of the  $N^2 - N$  incorrect pairs (Radford et al., 2021). This is done by maximizing the alignment between both image-text pair, pulling their embedding closer, versus random pairs, pushing their embedding farther in the embedding space. This loss consist of maximizing the posterior probabilities of image embedding given its corresponding text embedding and the other way around, this way it ensures that the image-text correlation is asymmetric to either modality.

The loss for the image encoder can be denoted as in Equation 7, where as the loss for the text encoder can be denoted as in Equation 8.

$$L_i(U, V) = -\frac{1}{n} \sum_{u_i \in U} \log \left( \frac{\exp\left(\frac{v_i^T u_i}{\tau}\right)}{\sum_{v_j \in V} \exp\left(\frac{u_i^T v_j}{\tau}\right)} \right) \quad (7)$$

<sup>1</sup><https://pubmed.ncbi.nlm.nih.gov/>

<sup>2</sup><https://www.wikipedia.org/>

<sup>3</sup><https://huggingface.co/>

$$L_T(U, V) = -\frac{1}{n} \sum_{v_i \in V} \log \left( \frac{\exp\left(\frac{u_i^T v_i}{\tau}\right)}{\sum_{u_j \in U} \exp\left(\frac{v_i^T u_j}{\tau}\right)} \right) \quad (8)$$

where  $\tau$  is a learnable temperature to scale logits, and it is fixed to 0.07. It controls the range of the logits and is directly optimized during training as a log-parameterized multiplicative scalar to avoid turning as a hyper-parameter (Radford et al., 2021). The similarity between the projected image embedding  $v_i$  and text embedding  $u_i$  is measured by the dot product between the embeddings.

The overall loss for a batch of image or exam and text pairs using  $U, V$  notations can be described as the average of  $L_I$  and  $L_T$  as in Equation 9.

$$L_{\text{CLIP}}(U, V) = \frac{1}{2} (L_I + L_T) \quad (9)$$

### 3.4. Interpreting Model Predictions and Outputs

At prediction, our network outputs logits, which are unnormalized predictions, for each input text prompt as shown in Figure 2b. We normalized the logits to obtain normalized probabilities using a *softmax* layer, and thus we match the text prompt with the highest similarity as the correct prediction to the input image or exam. Figure 3 shows different evaluation examples we generated on different classification tasks using the same input image and different input text.

### 3.5. Evaluation Procedure

We evaluated our implementation based on the experiments defined in Table 1, using both supervised classification and zero-shot classifications settings. The objective of comparing our image-label model trained on malignancy classification to the same encoder used in the network, which is a CNN, was to ensure that the model is able to perform an easy binary or multi-class classification task, thus we evaluated it using supervised approach. We reported the Binary Area Under ROC (AUROC) curve for binary tasks, and average AUROC with standard deviation for multi-class tasks.

We then added more complexity in terms of visual information or textual information (generated prompts sentences or reports or both combined) and measured the performance using zero-shot classifications using a class-specific generated prompts, as demonstrated in Figure 2b. We performed bootstrapping on 1000 samples, and averaged the AUROC of all of them, with the 95% confidence interval for binary tasks, and average AUROC with standard deviation for multi-class tasks. We also performed data-efficiency evaluation on different training data percentages for zero-shot evaluation. All experiments that uses single image as input will be referred to as "image level", whereas all experiments

that uses an exam with several images will be referred to as "exam level".

We also demonstrated the benefit of utilizing projection layers on top of the encoders we used by plotting t-SNE by Van der Maaten and Hinton (2008) of the image embeddings.

### 3.6. Computational Resources

All experiments were conducted on a NVIDIA TITAN V GPU with 12GB of memory. The code was implemented using PyTorch 1.13.1+cu116 in a Linux environment.

## 4. Experiments Results and Discussion

### 4.1. Datasets

**Image-Label** dataset is annotated at the image level, consisting of one mammogram view and several annotation labels. At the high level, it consisted of 3311 benign annotated files, and 3174 annotated as soft tissue lesions (STL) files, making a total of 6485 samples. Those files contained other several region level annotations, such as architectural distortion, benign or malignancy, calcification cluster or mass, and properties such as histology, mass shape, mass margin, mass density, and subtlety.

Among all of the samples, we re-split the dataset into more image level labels, either benign or malignant. Those image views that were known as malignant, but has benign label were eliminated as they could be wrongly labelled. Thus a total of 3311 benign samples, and 1653 malignant samples, with their internal region level annotations. Table 2 summarises all of the labels we used from this annotated dataset. Any "unknown" label within this table means that the label was missing in the original dataset.

Another internal annotated dataset that was used consisted of 9696 ground truth annotations for other image views samples (or included). This dataset consisted of several annotations such as malignancy, asymmetry, calcification, mass, histology, biopsy and several others.

**Exam-Reports** is an internal dataset that contains four image views per exam (or less views if they were not collected or available), and a long Dutch report. It consists of 10,801 exam-report samples. Among all of those samples, only 1832 were applicable to be used, excluding several pathology, biopsy, or duplicates and only selecting mammogram reports. We also extracted labels from the sentences and manually translated them to their English labels found in BI-RADS guidelines to minimize the translation error.

**Multi-label Prompts** are sentences generated randomly that contain one or more labels information. These sentences are formed by randomly selecting a template sentence describing each label, and concatenating them to form one or more sentences describing



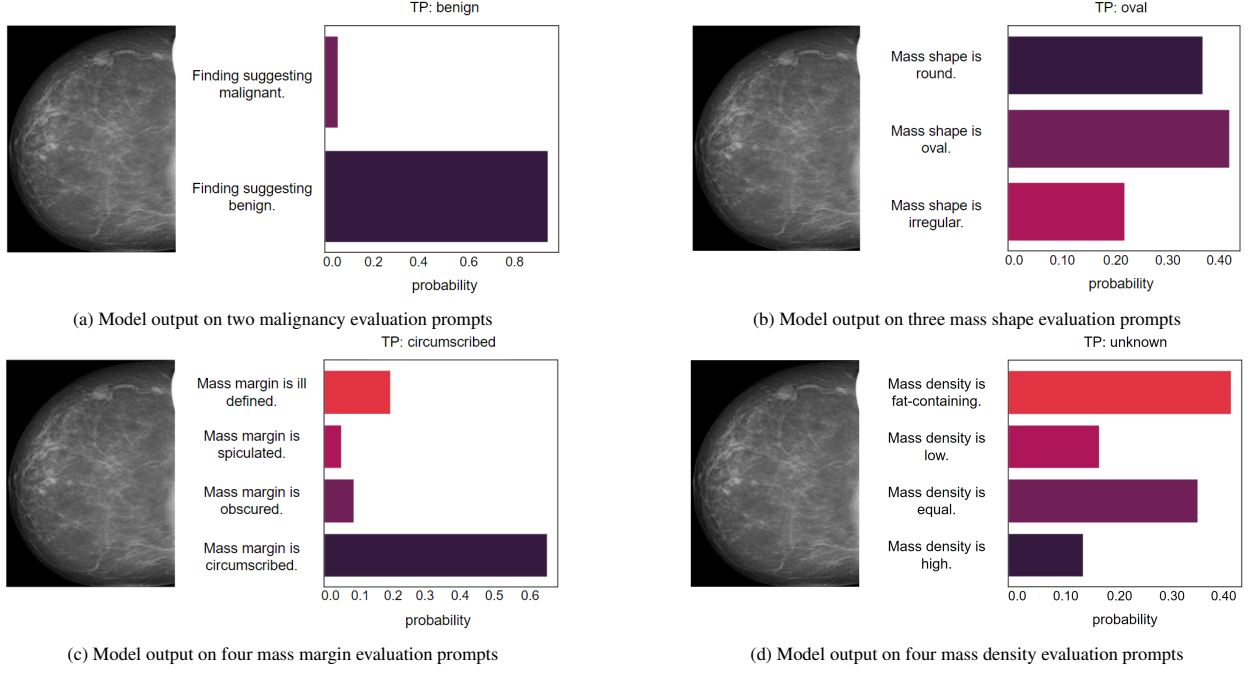


Figure 3: Demonstration model inference output on four different examples, where all of the output similarities are normalized. We run different inference text prompts on the same input image. In the figures, *TP* stands for True Positive.

the image or exam. Thus, forming a structured paragraph used to train the network. The labels used for generating the prompts are from any of the labelled datasets, and the additional labels extracted from the reports. The prompts text can be paired to either image or exam level datasets, as explained in Table 1. The process of generating the prompts can be found in Appendix B.

Table 3 summarizes the split of the datasets used for training, validation, and testing, where it was (70%, 15%, and 15%) respectively. To make the results comparable, the exam-reports dataset test split was the exact same test split for the image-label datasets.

#### 4.2. Baseline

**ConvNext Tiny** model (Liu et al., 2022), that is the same model used as an image encoder in our approach. This encoder will be used as the baseline for malignancy detection, when comparing to our models trained on image-label experiment dataset.

#### 4.3. Implementation Details

For the visual information, both at image and exam levels, we did not perform any augmentation or pre-processing. As text reports were originally in Dutch language, we translated them after pre-processing to standardise the training in English using the command `=GOOGLETRANSLATE(text_column, "nl", "en")` in Google Spread Sheets<sup>4</sup>. Pre-processing included eliminating unnecessary reports samples, text cleanup that

includes cleaning redundant words, structures, spaces, special characters, or patterns. As the nature of the mammography reports could include additional gold standard information that assist in evaluation of abnormalities, such as current study, ultrasound, mammogram X-ray, MRI, pathology, we selected only three types that we found contains most of the important information, that are current study, mammogram X-ray, and MRI. This was also performed during the pre-processing. The post-processing of the text was performed after the translation mainly to remove any duplicate sentences within the text, as the performance will heavily rely on the translation performance.

As for the embeddings, the final image and text embedding sizes are 512. Both encoders were frozen and only linear layers were trained on top of them. For both image-label (either binary or multi-class) and image-prompts experiments training, we used a single linear layer. For any of the exam level experiments, we used a 2 trainable linear layers with a ReLU activation function and dropout layer. By experimenting, we used dropout value of 0.2. For the training, we tracked the validation loss curves and several other area under the ROC (AUROC) values.

For all of the experiments, the early stopping condition was set with patience of 5 monitoring the validation loss and a tokenizer sequence length of 256. For the hyper-parameters, we used a cosine-annealing learning-rate scheduler (Loshchilov and Hutter, 2016), with a warm-up epoch of 0.1 and 30 trainable epochs, AdamW (Loshchilov and Hutter, 2017) optimizer with an initial learning rate  $5e-5$ , [EOS] token’s final output as the

<sup>4</sup><https://www.google.com/sheets/about/>

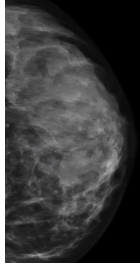
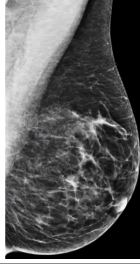
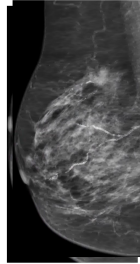
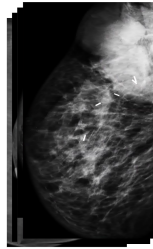
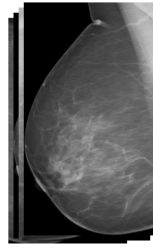
Experiment Name	Description	Input Example
Image-Label	Training with images and labels.	 <p>"benign"</p>
Image-Prompts	Training with images and prompts generated.	 <p>"Imaging revealed a mass with spiculated margins and irregular shape, suggestive of malignant pathology."</p>
Exam-Reports	Training with exams and reports text.	 <p>"Status after amputation of left breast due to carcinoma. Palpable abnormality on the right at 10 o'clock of 1.5 cm with skin retraction....."</p>
Exam-Reports + Prompts	Training with exams and reports text combined with prompts.	 <p>"The mass was characterized by ill defined margins and oval shape on imaging, suggesting a potential malignant etiology, assigning BIRADS score of 4 based on the findings. Fast growing tumor of right breast/axilla. Excision in February followed by recurrence. (PA of excision inconclusive). Currently malignant ....."</p>
Exam-Prompts	Training with exams and prompts generated.	 <p>"The mass displayed spiculated margins, suggestive of a malignant lesion, the mammography report assigns a BIRADS score of 5 to guide further clinical decisions."</p>

Table 1: Experiments description and the datasets used in each of them.

global textual representation, and weight decay  $1e-4$  following the work of You et al. (2023). For image-label experiments, we used a batch size of 32 samples for all three splits, whereas for the remaining experiments, we used batch size of 64.

#### 4.4. Classification

We started by evaluating the learned representation on several image classification tasks based on our image-label dataset available labels mentioned in Table

2, using both supervised image classification and zero-shot classification settings. In both settings, as mentioned earlier, we only trained linear projection layers on top of the pre-trained encoders.

##### 4.4.1. Supervised Image Classification

For the supervised classification, as our baseline CNN encoder was pre-trained on malignancy task, we trained our network on the malignancy labels of the image-label dataset, and compared the results area un-



Label Group	Labels Names	Count
Malignancy	Benign	3311
	Malignant	1653
Mass Margins	Unknown	2467
	Ill defined	1095
	Obscured	697
	Spiculated	484
	Circumscribed	221
Mass Shapes	Unknown	2466
	Irregular	1218
	Round	681
	Oval	599
Architectural Distortion	Normal	4842
	Distortion	122
Calcification	No Calcification	2969
	Has Calcification	1995
Mass	No Mass	278
	Mass	4686

Table 2: Image-Label dataset description.

Dataset	Split	Count
Image-Label <i>or</i> Image-Prompts	Train	3474
	Valid	1490
	Test	745
Exam-Reports <i>or</i> Exam-Reports + Prompts <i>or</i> Exam-Prompts	Train	1282
	Valid	550
	Test	745

Table 3: Datasets split summary. First row summarizes the image level splits, either using labels or prompts depending on the experiment, and second row summarizes the exam level splits.

der the ROC curve (AUROC) of the true class. We also trained a network for the other labels of the dataset and reported the results in Table 4. In our results, we show that our network was able to outperform a traditional CNN performance on malignancy detection by training a single linear layer. Our network also performed well on the remaining classification tasks. The main objective was to ensure that the network is capable of learning a simple label classification task, either binary or multi-class using the learned representation from both image and text modalities.

#### 4.4.2. Zero-shot classification

For the zero-shot prompt classification, the network was trained and evaluated on different experiments, thus different representations. The constructed evaluation text prompts were specified to target the model performance in understanding the clinical meaning of the text input as a full sentence. Therefore, we constructed a class-wise inference prompt for each label task. Those inference prompts are different from the prompts generated for training, and can be found in Table 6. We evaluate the binary classification tasks by computing the AUROC of 1000 bootstrapped samples with 95% CI, and

computed the average AUROC for multi-class classification tasks with standard deviation. We also evaluated the performance on both datasets, at image and exam level training, and to make the evaluation fair, all experiments were evaluated on the same test samples at the image level.

As shown in Table 5, both experiments image-prompts and exam-prompts outperform all other experiments, where those experiments were trained on different dataset samples, and on the same text prompting approach we proposed. Training the network with well structured sentences as the generated prompts performs better than training with real radiologist reports as the nature of the text reports when they are written, they are not generally standardised. This can be also demonstrated when training the network with exam-reports and exam-reports + prompts, where including the prompts improved the results as demonstrated in the table. It is also worth noting that each experiment row in Table 5 is a single model performance, thus shows the ability in generalizing to different downstream tasks.

#### 4.5. Data-efficiency Evaluation

We further evaluated the model performance for zero-shot classification taking into account different sizes of training dataset samples (10%, 20%, 50%, and 100%), on malignancy detection. In Figure 4, we show that both of our models, either trained on image-label malignancy task, or on exam-prompts experiments improve the performance when more training data is used, tracking their malignancy AUROC metric for all of the test samples. The image-label trained model shows only slight improvement as the encoder only performance (in red color) is high, so training linear layers on top of the pre-trained encoder improves its ability in malignancy zero-shot classification for this specific dataset. It demonstrated a consistent high performance on all percentages of the training data. The exam-prompts model that is trained on more visual and textual information showed a significant improvement in the malignancy zero-shot detection with different percentages, indicating that the model is effectively learning from the additional data.

#### 4.6. Report Generation

To generate a radiology report, we defined a report as a series of zero-shot classification tasks. Those can be specific based on BI-RADS mammography guidelines, or general to any other inference task. To generate a report, we used the exam-prompts experiment model, and constructed a series of inference tasks. The final step of the report generation includes formatting all outputs into a template sentences and concatenating the results to form a single report. In Figure 5, we demonstrate a summary of our report generation pipeline. At the top level, an inference task is made to validate if an image

Experiments	Binary AUROC $\uparrow$				Average Multi AUROC ( $\pm$ std) $\uparrow$	
	Malignancy	Arch. Dist.	Mass	Calcification	Mass Shapes	Mass Margins
CNN (Baseline)	0.9153	-	-	-	-	-
Image-Label	0.9402	0.8293	0.8005	0.8820	0.8023 ( $\pm$ 0.078)	0.8344 ( $\pm$ 0.089)

Table 4: Comparison of area under the ROC (AUROC) of different experiments and classification tasks (binary and multi-class) using one-vs-all classification evaluation on image level experiments. Total 745 samples of the image-label dataset test split were used. In the table headers, Arch. Dist. stands for architectural distortion.

Experiments	Average Binary Bootstrap Samples AUROC (95% CI) $\uparrow$				Average Multi AUROC ( $\pm$ std) $\uparrow$	
	Malignancy	Arch. Dist.	Mass	Calcification	Mass Shapes	Mass Margins
Image-Prompts	<b>0.931</b> <b>(0.905-0.953)</b>	0.682 (0.554-0.808)	0.663 (0.564-0.755)	0.680 (0.639-0.719)	0.727 ( $\pm$ 0.120)	<b>0.715</b> <b>(<math>\pm</math> 0.154)</b>
Exam-Reports	0.828 (0.791-0.861)	0.637 (0.504-0.78)	0.475 (0.3721-0.572)	0.567 (0.524-0.610)	0.596 ( $\pm$ 0.079)	0.560 ( $\pm$ 0.089)
Exam-Reports + Prompts	0.847 (0.814-0.878)	0.646 (0.509-0.791)	0.527 (0.425-0.619)	0.683 (0.644-0.723)	<b>0.848</b> <b>(<math>\pm</math> 0.088)</b>	0.594 ( $\pm$ 0.094)
Exam-Prompts	0.916 (0.891-0.938)	<b>0.717</b> <b>(0.620-0.804)</b>	<b>0.678</b> <b>(0.603-0.743)</b>	<b>0.736</b> <b>(0.701-0.772)</b>	0.700 ( $\pm$ 0.106)	0.639 ( $\pm$ 0.218)

Table 5: Comparison of the average area under the ROC (AUROC) of different experiments and classification tasks (binary and multi-class) using zero-shot classification evaluation on both image and exam level experiments. For binary tasks, we bootstrapped 1000 samples, and computed the average AUROC and 95% CI. For the multi-class tasks, we computed the average AUROC  $\pm$  standard deviation. Total 745 samples of the image-label dataset test split were used. In the table headers, Arch. Dist. stands for architectural distortion.

Label Group	Input Evaluation Prompt
Malignancy	Findings suggesting {label}.
Mass Margins	Mass margins is {label}.
Mass Shapes	Mass shape is {label}.
Architectural Distortion	Normal architecture is visible. Displayed architectural distortion.
Calcification	No calcifications are present. Finding suggesting calcifications.
Mass	No mass was observed. Findings revealed a mass.

Table 6: Zero-shot evaluation prompts for all label groups. The {label} are replaced with the labels reported in Table 2, that are based on BI-RADS guidelines.

or exam either has a mass, calcification, or no findings. As “No Findings” ends the report, it doesn’t require any further evaluation for mass or calcification information, thus we report directly a conclusion sentence as shown in Figure 6b.

Both “Mass” and “Calcification” in Figure 5 have their own generation pipeline. In “Mass” track, we evaluate the malignancy, mass shape, mass margins, BI-RADS score, and architectural distortion. As for “Calcification” track, we evaluate malignancy, distribution, BI-RADS score, and architectural distortion. An example for a report generated for an exam with malignant mass findings is shown in Figure 6a, where as an ex-

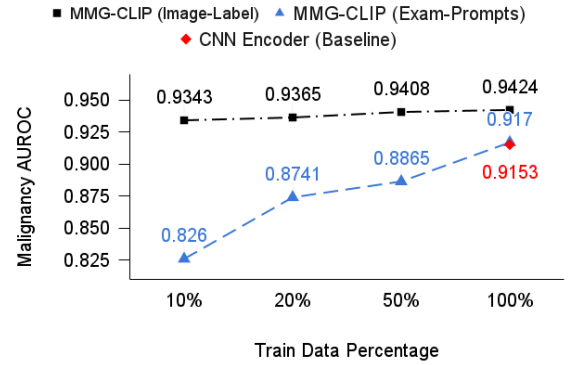


Figure 4: Image-label and exam-reports models (ours) zero-shot performance for malignancy classification using different amount of data, without bootstrapping.

ample for a report generated for an exam with benign calcification is shown in Figure 6c.

One important limitation of our report generation is the decision condition taken for all prompts output similarities generated from the model. If the model fails on identifying the correct type of findings at the very first level of the generation pipeline, all following evaluation results will be wrong. Figure 6d shows a failed example of a report generated as “No Findings”, where it contains other types of findings. As we take the maximum similarity value of all text-prompts output similarities,

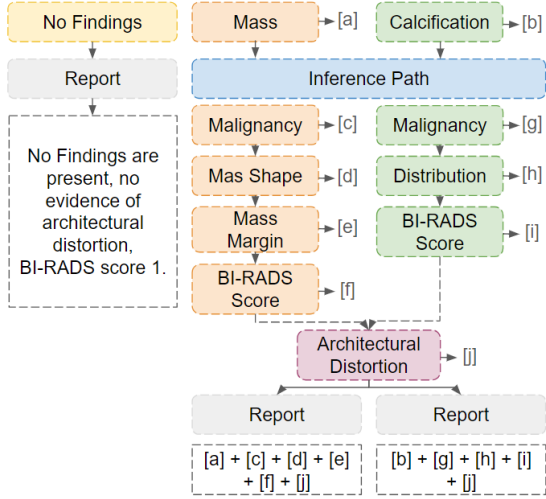


Figure 5: Report generation pipeline. Symbol  $[letter]$  represent the inference task output, and  $+$  represent output formatting and concatenation.

we are not able to distinguish between a strong prediction (with high probability for a specific text prompt) or for a confused prediction (when all probabilities are close to each others). Another concern is whether an exam or an image has more than one finding similar to both “Mass” and “Calcification” together. When taking the maximum similarity, we result with having only one output text to the inference task, thus can’t combine multiple texts as an output.

#### 4.7. Embedding Visualization

Data visualization using dimension reduction approaches can assist in understanding the geometric and neighborhood structures of datasets (Wang et al., 2021). A popular tool to perform dimensional reduction is the t-distributed Stochastic Neighborhood Embedding (t-SNE) algorithm (Shah and Silwal, 2019), introduced by Van der Maaten and Hinton (2008), or principal component analysis (PCA). We performed t-SNE analysis on both the embeddings from the CNN encoder as in Figure 7a and 7c, as well as on the linear layers on top of the encoder as in our model in Figures 7b and 7d. The figures demonstrates the separation of “benign” and “malignant” embeddings classes of the malignancy classification task as an output of the networks projected into lower dimensional using t-SNE.

As shown in Figures 7a and 7b, both networks generates a well clustered points of both labels. Using the CNN only however shows some overlap between the two classes, indicating that the baseline CNN encoder does not completely distinguish between them. Adding linear layers on top of the pre-trained encoder does slightly produce better clusters as it focuses on the specific characteristics and patterns present in our datasets, thus making it perform better on our test cases and provides more distinct clusters with less overlap. We also

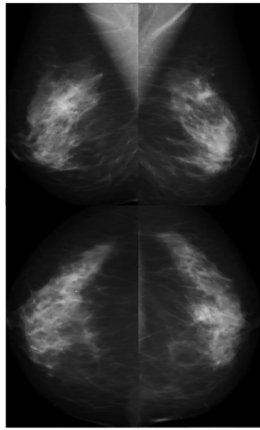
visualized the first dimension of t-SNE with respect to the models probabilities to belong to malignancy class as shown in Figures 7c and 7d. Both Figures indicates positive correlation between the t-SNE dimension 1 and malignancy probabilities, where the model with projection layers as in Figure 7d shows more distinct and reliable probability estimates for malignancy, as there is a clearer separation between both labels cases compared to the baseline encoder alone in 7c.

#### 4.8. Limitations and Future Work

Despite that we reached promising results in our experiments, we believe that there are improvements that can be made.

**Embedding pairing** is a challenging task in medical image-text datasets as the nature of the visual and textual information can be paired to more than one sample. For example, an image can contain several regions of interest, where it can be described correctly in two separate reports sentences of two different exams. This makes the loss metrics not meaningful when it comes to training as pairing a single image-text pairs might not be meaningful when the network learns the global representation of all of the input data. This also was observed when training with large batches (that are possible to have reports with similar information) on a small datasets like ours, but not observed when using a very small batch size as the possibility of having two samples of same findings is much lower. We experimented implementing different variations of CLIP InfoNCE loss taking into account the batch samples and other sampling mechanisms to tackle the problem, however none of the approaches we tried proved better learning when it comes to long text reports. Thus, a meaningful contrastive loss would be very beneficial for the network to be able to match medical image-text datasets. For example, region-wise matching between image and text information, or giving more weights to certain regions could potentially improve the network loss mechanism.

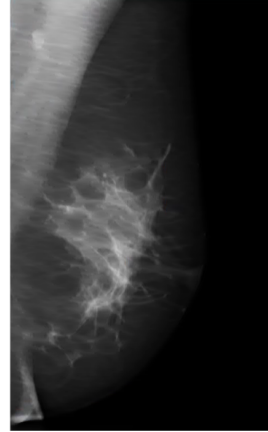
**Report generation.** When it comes to generating report, as we mentioned earlier we use the maximum similarity output as the final task result before creating a report. Trying different decision making approaches could be useful in generating more precise reports, but it also requires human intelligence and clinical validation. One case that we noticed could be failing repetitively is when network received 5 input prompts, and the five similarities values are very close to each other, using the maximum value might not be ideal. Also, some report details have more importance than others, for example malignancy classification, or differentiating between the presence of mass, calcification, or no findings, compared to other sub-tasks like mass region or calcification distribution. The decision making here plays an important role in the report accuracy, and taking the maximum similarity value might not always be the best case. Thus, other decision making approaches such as



Generated Report:

*"The mass demonstrated spiculated margins and irregular shape, prompting further evaluation for malignant features, this concludes assigning a BIRADS score of 0. No evidence of architectural distortion was noted on mammography."*

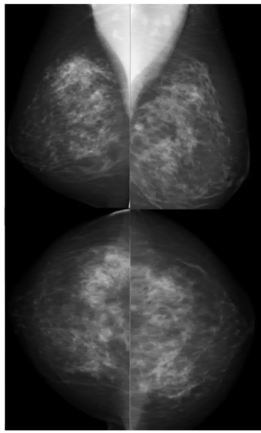
(a) Exam level generated report revealing a malignant mass



Generated Report:

*"No findings are present. Mammography showed no evidence of architectural distortion. BI-RADS score 1."*

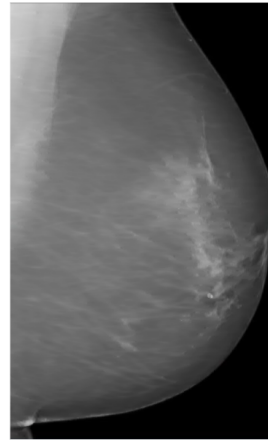
(b) Image level generated report revealing no findings



Generated Report:

*"Observed calcifications appear benign with regional distribution, assigned BIRADS score 3 for clinical management. The presence of architectural distortion on mammography necessitated careful evaluation."*

(c) Exam level generated report revealing a benign calcification



Generated Report:

*"No findings are present. Mammography showed no evidence of architectural distortion. BI-RADS score 1."*

(d) Image level failed generated report

Figure 6: Demonstration of report generation using a full exam as in (a), and (c), or a single image view as in (b) and (d). Text highlighted in green is a correct prediction from the network, where text highlighted in red are wrong predictions. Yellow highlighted text has no label to compare with.

applying a threshold value to the similarities could be explored in future work.

**Pre-training** the encoders on large scale datasets could significantly improve the performance when and generalization of the model. As we used pre-trained encoders, the extracted features relies on their performance as well as on the performance of the trained linear layers. And as we had a very small amount of data to work on, we were not able to train the models from scratch.

**Network Architecture** can be improved to localize the presence of the pathology reported in the text to which exam view it is found in. This can significantly improve the reporting precision if the network is capable of identifying which view exactly has more importance. In our implementation, while training the network, we averaged the features extracted from each of the input image views, thus losing the anatomical location of the pathology it contains. For example, when a mass appears in the "right MLO image view" in an exam, we lose such information while averaging the embedding. Having that considered can also improve the feature ex-

traction approach to assign more weights to important views and less to others.

## 5. Ablation Study

**Ablation on model architecture.** To understand the effectiveness of the architectural parameters and key components, we conducted ablation study using different parameters and components with respect to malignancy zero-shot classification performance. All results reported in Table 7 were trained using the best exam-prompts experiment model. We used different training configurations to evaluate their impact on zero-shot classification performance on one task.

In the first row, we evaluate different number of projection layers. From the reported results, 2 Linear Projection layers gave the best zero-shot performance for our model and no indication of increased performance when more trainable layers are used.

In the second row of Table 7, we used the default 2 projection layers with different training batch sizes. The default value we used was batch size  $n=64$  with

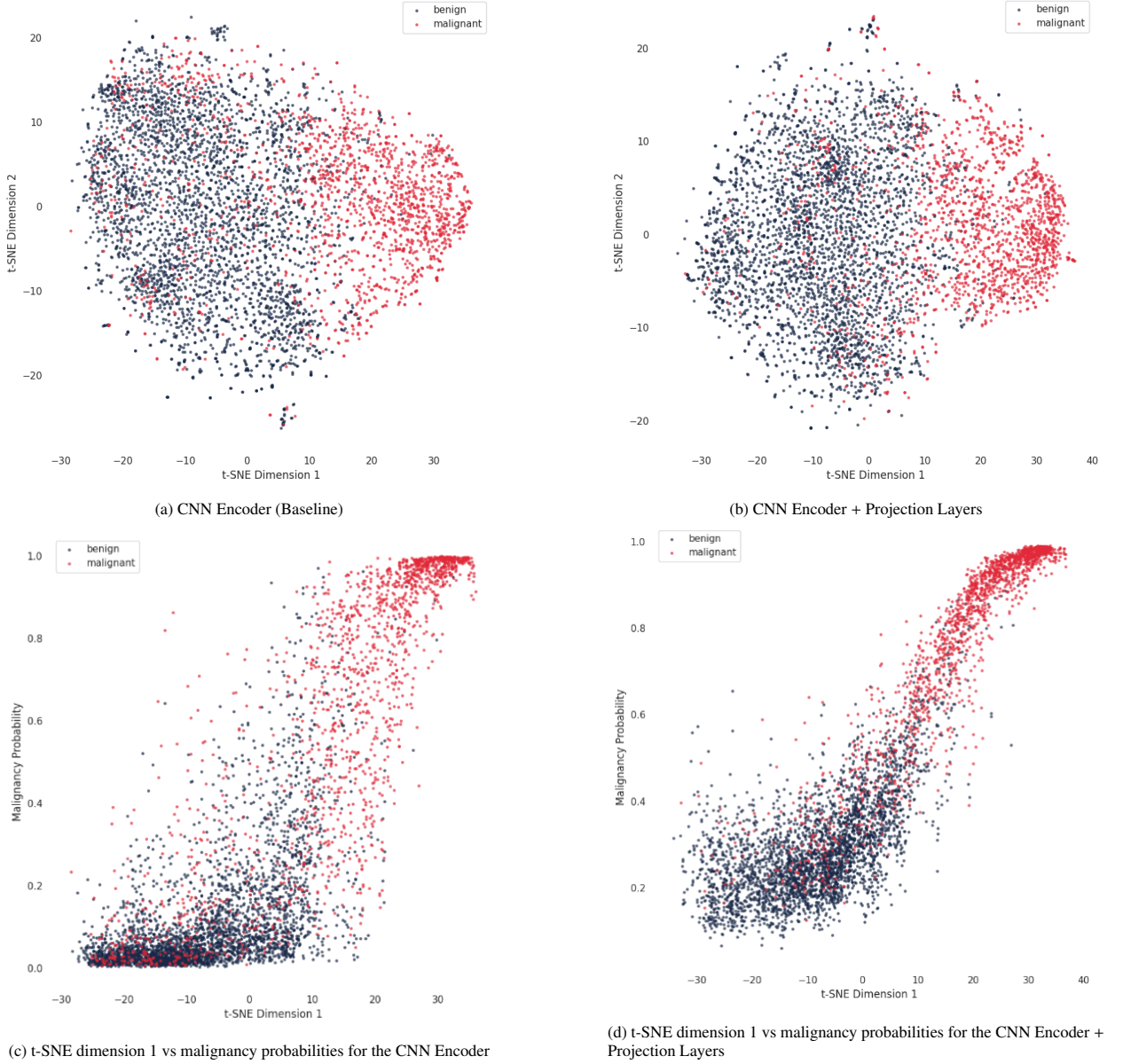


Figure 7: Image embedding visualization of malignancy Image-Label dataset for both CNN encoder (Baseline) alone ours that includes projection layers on-top of the encoders.

tokenizer sequence length of 256 for exam-prompts experiment model where it obtained 0.916 (0.891-0.938). Both  $n=32$  and  $n=128$  showed no significant improvement on the performance as reported in the table. Similarly to the tokenizer sequence length in the third row, both sequence lengths 384 and 512 didn't improve the performance of our default value. In addition to that, we observed that using a logit scale  $\tau = 0.07$  performs better than without performing scaling to the logits during training as in the last row reported in the table.

**Ablation on inference prompts.** As mentioned previously, our evaluation prompts contribute significantly to the results we obtained, as we believe it targets the clinical meaning behind the label we are evaluating. To measure the impact of changing the inference prompts

during zero-shot settings, we experimented using CXR-CLIP by You et al. (2023) evaluation prompts for zero-shot and compared the results to ours. In this evaluation, we are not comparing our results to theirs, as it is using totally different datasets in different domains, but only comparing our model behaviour to different evaluation prompts. In CXR-CLIP, they used the prompts “{classname}” versus “No {classname}” for all labels they evaluate, for example “No oval” versus “oval” for “Mass Shapes” task, and then using prediction of the “{classname}” to generate the results. We noticed that this introduces a challenge for our network when it comes to multi-class evaluation, where it performs poorly using their prompting mechanism compared to ours, for example “Mass shape is oval”. Table 8 shows



Experiments	AUROC (95% CI) $\uparrow$
MMG-CLIP	
w/ 1 proj. layers	0.893 (0.864-0.920)
<b>w/ 2 proj. layers</b>	<b>0.916 (0.891-0.938)<sup>a</sup></b>
w/ 3 proj. layers	0.910 (0.882-0.933)
MMG-CLIP	
w/ batch size = 32	0.908 (0.883-0.933)
w/ batch size = 128	0.912 (0.885-0.936)
MMG-CLIP	
w/ seq. length = 384	0.910 (0.885-0.933)
w/ seq. length = 512	0.906 (0.877-0.929)
MMG-CLIP	
w/ logit scale = 1	0.8876 (0.858-0.913)
(no scale)	

<sup>a</sup> Value obtained using the default experiment parameters as 2 proj. layers, batch size = 64, seq. length = 256, logit scale  $\tau = 0.07$ .

Table 7: Ablation study of key architectural parameters with respect to different parameters and components. The reported scores are the average AUROC of 1000 bootstrapped samples with 95% CI on malignancy zero-shot classification. In the table, *proj. layers* is projection layers, *seq. length* is the tokenizer sequence length.

Experiments	AUROC ( $\pm$ std) $\uparrow$
MMG-CLIP	
w/ CXR-CLIP prompts	0.587 ( $\pm$ 0.074)
<b>w/ our prompts</b>	<b>0.700 (<math>\pm</math> 0.106)</b>

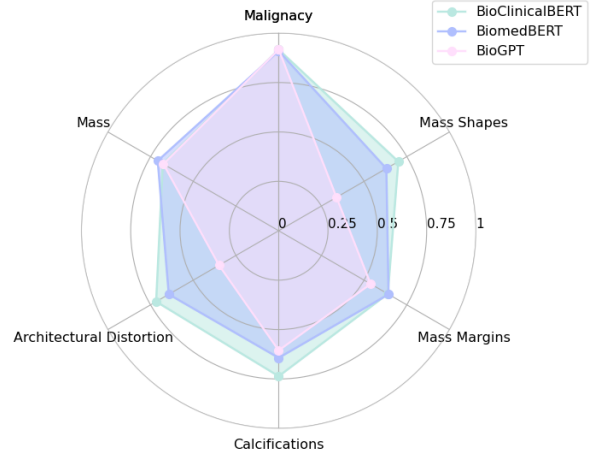
Table 8: Ablation study of different evaluation prompts used to evaluate zero-shot settings. The reported scores are the average AUROC ( $\pm$  std) for all labels curves on “Mass Shapes” task.

that with our evaluation, we obtain higher score for “Mass Shapes” task when using our prompts compared to using CXR-CLIP evaluation prompts.

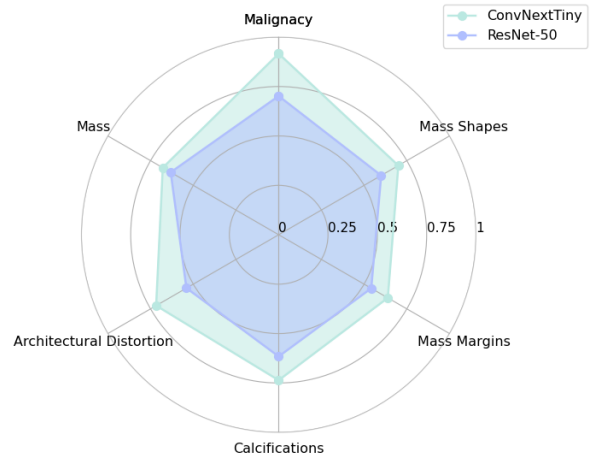
#### Ablation on pre-trained clinical text encoders.

As we used pre-trained text encoder BioClinicalBERT model by Alsentzer et al. (2019) and not pre-training our own due to the limited number of training data, the network performance heavily relies on the performance of the pre-trained text encoder. To understand the impact, we analyzed our network performance using other large language models of different parameter sizes as our text encoder. In Figure 8a, we compared the performance of our model trained using exam-prompts experiment, similar to the evaluation approach reported in Table 5.

As shown in Figure 8a, BioClinicalBERT model as our text encoder outperforms both BiomedBERT and BioGPT in performance for all zero-shot classification tasks. This supports idea of having a domain specific pre-trained model on clinical text datasets when it comes to learning medical text reports from other domains, and encourages pre-training a mammography specific text encoder for future work. Following BioClinicalBERT is BiomedBERT, where it shows a bal-



(a) Different LLM models performance as text encoders.



(b) Different vision models performance as image encoders.

Figure 8: Comparison between using different vision and large language models as encoders in our network on zero-shot classification tasks on the exam-prompts experiment model. Values on the axis (0, 0.25, 0.75, and 1) are average AUROC values for 1000 bootstrapped samples for binary tasks, and average AUROC for multi-class tasks.

anced performance across most tasks with particular strength for both “Mass” and “Malignancy”. It also supports the idea that BERT variant models tends to outperform GPT variant models which are more commonly used in generation tasks (Luo et al., 2022). BioGPT was the least in performance as it had very low metric values, close to randomness. Thus, both BioClinicalBERT and BiomedBERT are more suitable text encoders encoding medical text given their performance on our various tasks, and could potentially be used in pre-training a BERT model for mammography domain-specific data.

**Ablation on pre-trained vision image encoders.** To assess the performance of a domain-specific pre-trained model as an image encoder such as our pre-trained ConvNeXt Tiny image encoder, we used a ResNet-50 model and applied transfer learning approach to its last layer (layer 4), with similar training configurations. In Figure 8b, we show that having a pre-trained model on domain-



specific knowledge significantly outperforms a model pre-trained on general vision task, where our ConvNext Tiny model performed better in all tasks. ResNet-50 model had consistent and balanced performance and was not biased to a specific task.

## 6. Conclusions

In this work, we proposed an image-text contrastive learning framework named *MMG-CLIP* as well as a report generation BI-RADS specific pipeline for mammography X-ray 2D images. Our implementation includes not only training the network at the image or exam level (multiple images) with medical text, but also utilises multi-class generated prompt text to improve the model performance on zero-shot classification tasks. *MMG-CLIP* showcases remarkable flexibility due to the multi-modality and zero-shot learning ability. Our experiments results shows the network data-efficiency and zero-shot capability of the learned representations for various downstream classification tasks.

## Acknowledgments

I would like to express my deepest gratitude ScreenPoint Medical for providing me with this incredible opportunity to work on this thesis topic. First and foremost, I would like to thank my supervisors, Santiago Pires and Jaap Kroes for their guidance, support and mentoring. I'm also very grateful to all ScreenPoint teams for their assistant, feedbacks, and providing the necessary resources. Together, we achieved more than I could have ever accomplished alone. I would also like to extend my sincere appreciation to the MAIA master consortium for providing me with a solid educational foundation and to the European Commission for granting me this opportunity and generously funding my master education. Lastly, I would like to thank my family and friends for their endless support and unconditional love, despite the long distance.

## References

- Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M., 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.
- Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K., 2018. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems* 42, 1–13.
- Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M., 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* 66, 101797.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Fishman, M.D., Rehani, M.M., 2021. Monochromatic x-rays: The future of breast imaging. *European Journal of Radiology* 144, 109961.
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1–23.
- Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al., 2021. Pre-trained models: Past, present and future. *AI Open* 2, 225–250.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation* 9, 1735–1780.
- Huang, S.C., Shen, L., Lungren, M.P., Yeung, S., 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR. pp. 4904–4916.
- Jing, B., Xie, P., Xing, E., 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.
- Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S., 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* 6, 317.
- Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G., 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1–9.
- Kallenberg, M., Petersen, K., Nielsen, M., Ng, A.Y., Diao, P., Igel, C., Vachon, C.M., Holland, K., Winkel, R.R., Karssemeijer, N., et al., 2016. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE transactions on medical imaging* 35, 1322–1331.
- Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis* 65, 101759.
- Kisilev, P., Sason, E., Barkan, E., Hashoul, S., 2016. Medical image description using multi-task-loss cnn, in: Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1, Springer. pp. 121–129.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25.
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al., 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* 128, 1956–1981.
- Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J., 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.
- Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986.
- Loshchilov, I., Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y., 2022.

- Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* 23, bbac409.
- Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. *Journal of machine learning research* 9.
- Mohamed, A.A., Berg, W.A., Peng, H., Luo, Y., Jankowitz, R.C., Wu, S., 2018. A deep learning method for classifying mammographic breast density categories. *Medical physics* 45, 314–321.
- Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Pesapane, F., Trentin, C., Ferrari, F., Signorelli, G., Tantrige, P., Montesano, M., Cicala, C., Virgoli, R., D’Acquisto, S., Nicosia, L., et al., 2023. Deep learning performance for detection and classification of microcalcifications on mammography. *European Radiology Experimental* 7, 69.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: *International conference on machine learning*, PMLR. pp. 8748–8763.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 9.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28.
- Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports* 8, 4165.
- Salama, W.M., Aly, M.H., 2021. Deep learning in mammography images segmentation and classification: Automated cnn approach. *Alexandria Engineering Journal* 60, 4701–4709.
- Shah, R., Silwal, S., 2019. Using dimensionality reduction to optimize t-sne. *arXiv preprint arXiv:1912.01098*.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Wang, J., Yang, X., Cai, H., Tan, W., Jin, C., Li, L., 2016. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific reports* 6, 27327.
- Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M., 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058.
- Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y., 2021. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research* 22, 1–73.
- Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.
- Xie, X., Niu, J., Liu, X., Chen, Z., Tang, S., Yu, S., 2021. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis* 69, 101985.
- Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R., 2019. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 292, 60–66.
- You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B., 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 101–111.
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

## Appendix A. Data sampling for image-prompts experiment

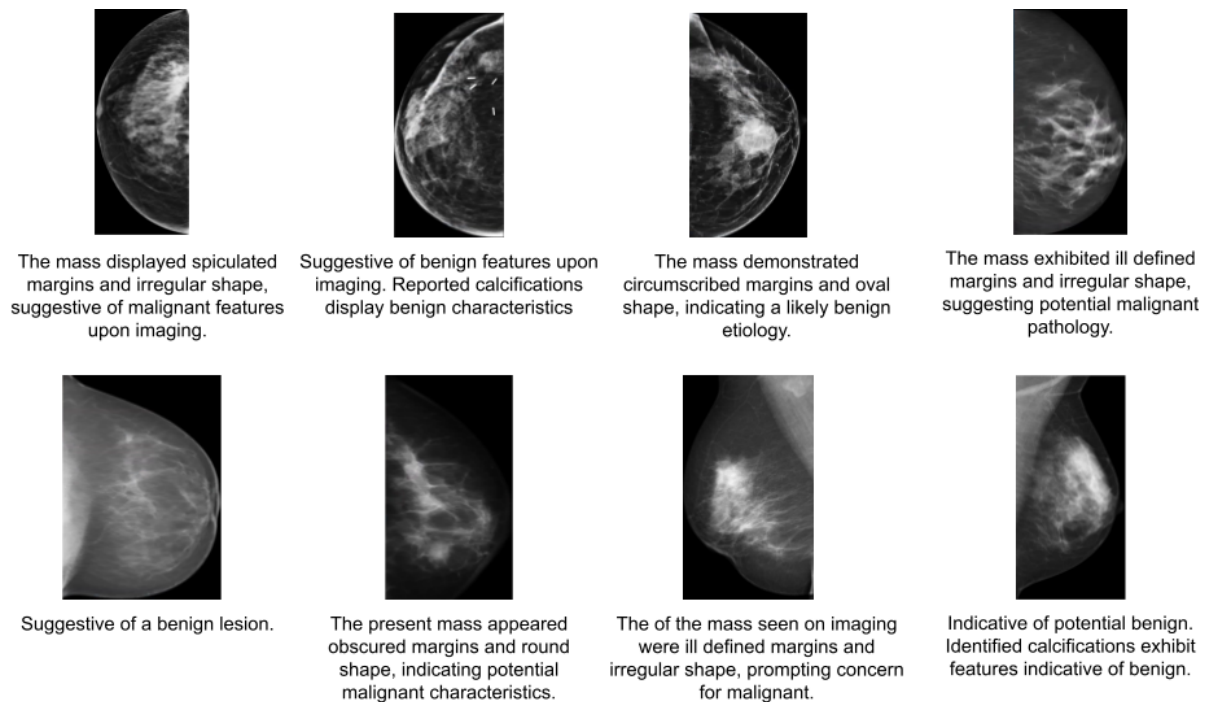


Figure .9: Example of image-prompts pairs sampled from training dataset.

## Appendix B. Prompt generation approach using labelled data.

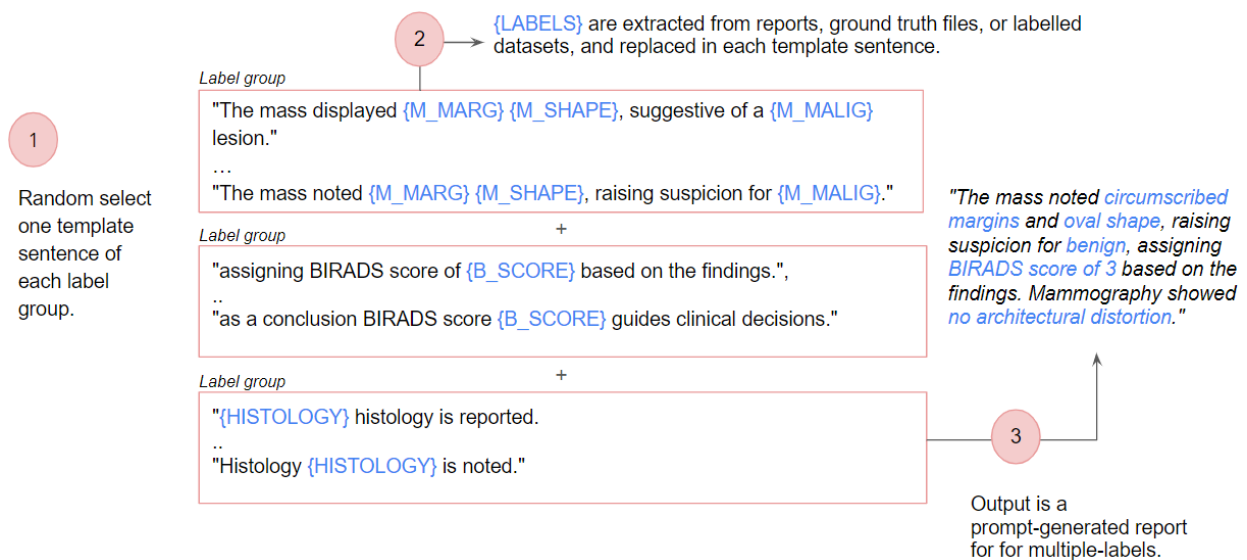


Figure .10: Demonstration of prompt generation mechanism using multiple labels.