SCREENPOINT
Medical

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

Presenter: Abdelrahman HABIB

transpara®
By ScreenPoint Medical

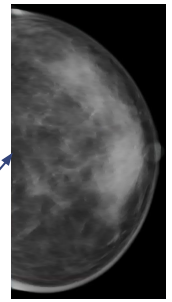
Table of Contents

Introduction

In this work, we tackle the task of automated mammography report generation following Breast Imaging Reporting & Data System (BI-RADS) guidelines.

Motivation

Lack of medical image w/ labels



Shape: Oval

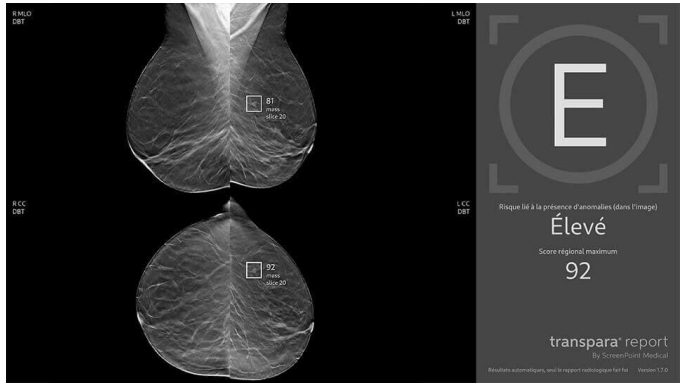
Shape: Round

Shape: Irregular

Utilize image radiology report text instead/with labels

The mass displayed spiculated margins and irregular shape, suggestive of a malignant lesion …

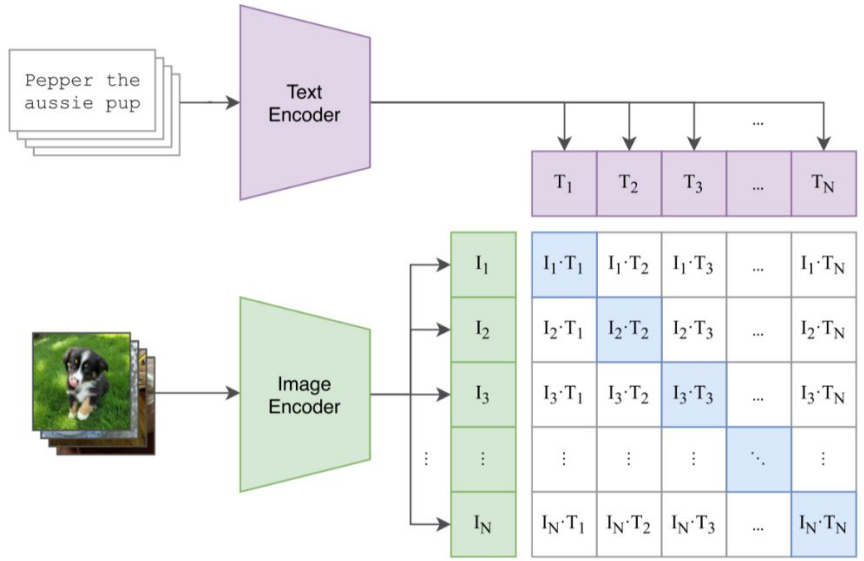
AI products lack the explainability for its decision, e.g. score is 92, but why did it make such decision?



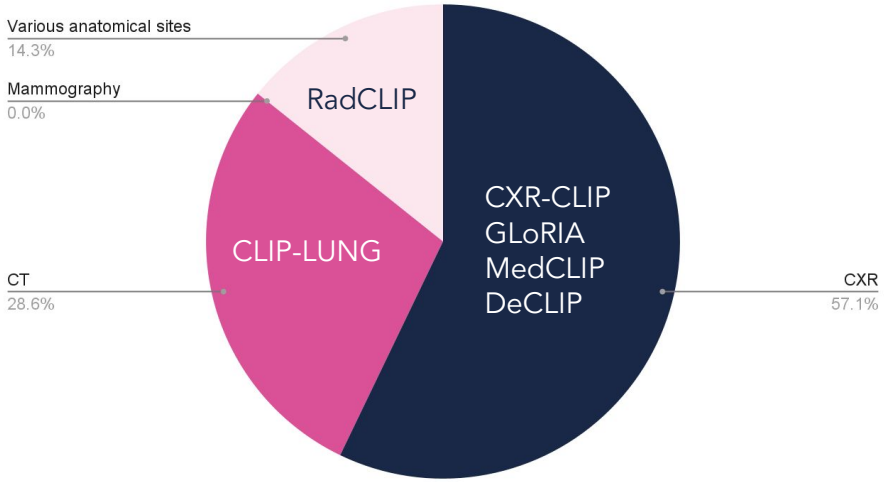
Radiologist prefer explanations & interpretations of such results, as text reports along with generated scores or annotations.

Research Background

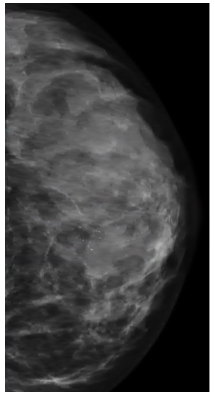
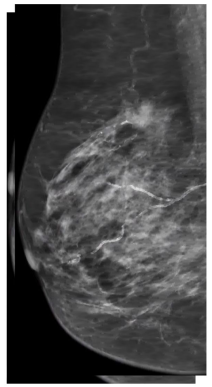
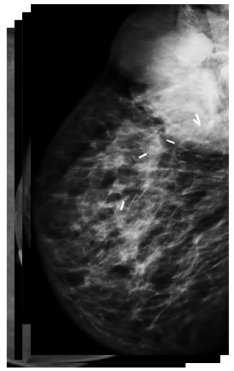
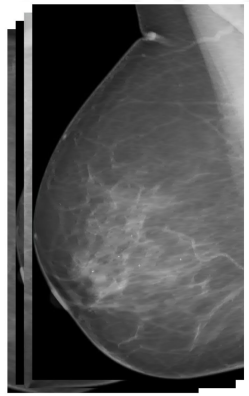
Contrastive Language-Image Pre-training (CLIP) allows training with image-text pairs by jointly training two encoders, and not relying on labels.



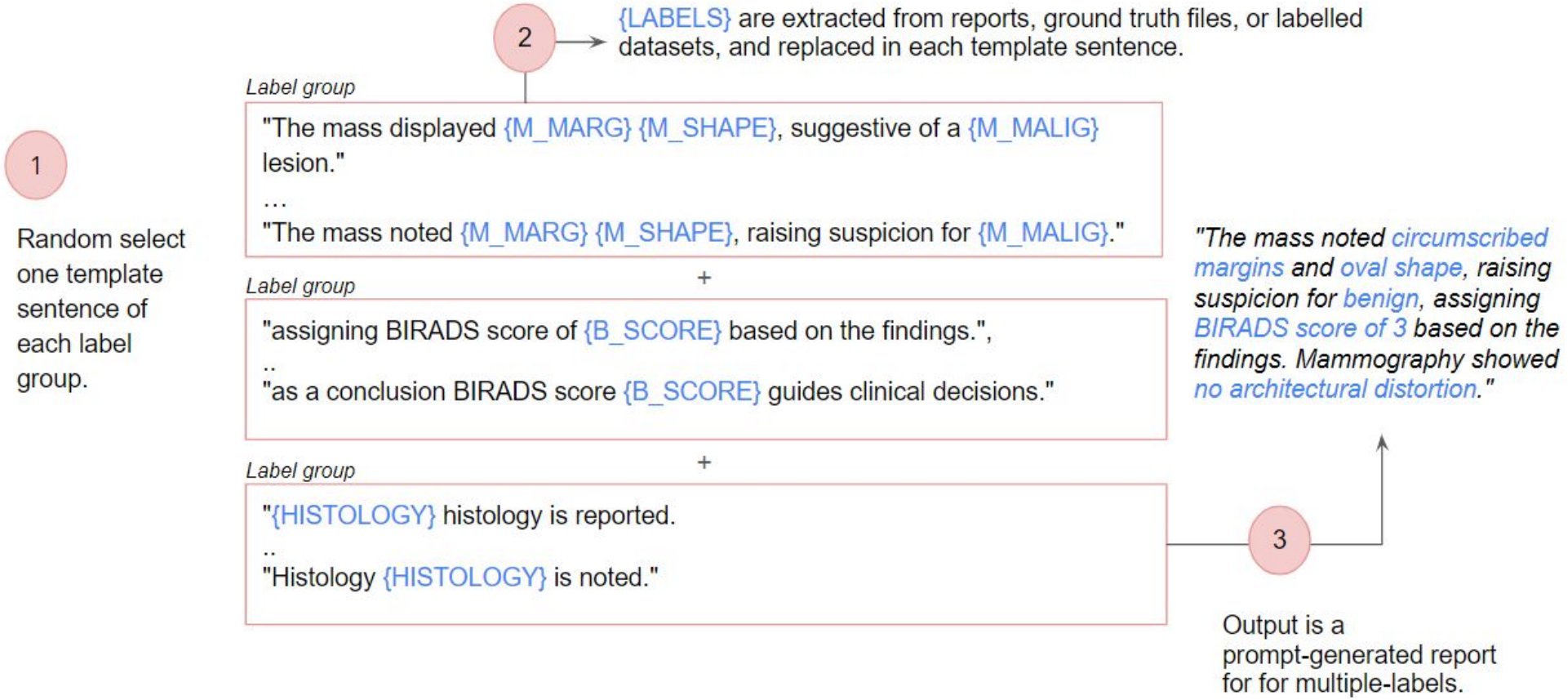
Most of the work that utilized CLIP was focused on different domains. No work was focused in mammography report generation.



Datasets & Experiments

Image-Label	Image-Prompts	Exam-Reports	Exam-Reports+Prompts	Exam-Prompts
 *"benign"*	 *"Imaging revealed a mass with spiculated margins and irregular shape, suggestive of malignant pathology."*	 *"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….."*	 *"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 ……"*	 *"The mass displayed spiculated margins, suggestive of a malignant lesion, the mammography report assigns a BIRADS score of 5 to guide further clinical decisions."*
Training with images and labels (for example 'benign' or 'malignant' label).	Training with images and a variation of generated text reports based on prompts templates.	Training with all images of an exam and radiology reports.	Training with all images from an exam, radiology reports and generated text reports.	Training with all images of an exam and generated text reports.

Generating text reports based on prompts templates (generated from annotations)



Our approach towards report generation

More Training Complexity

Train a network on image-label data, must perform as good as a CNN (binary & multi-class image-label classification).

↓

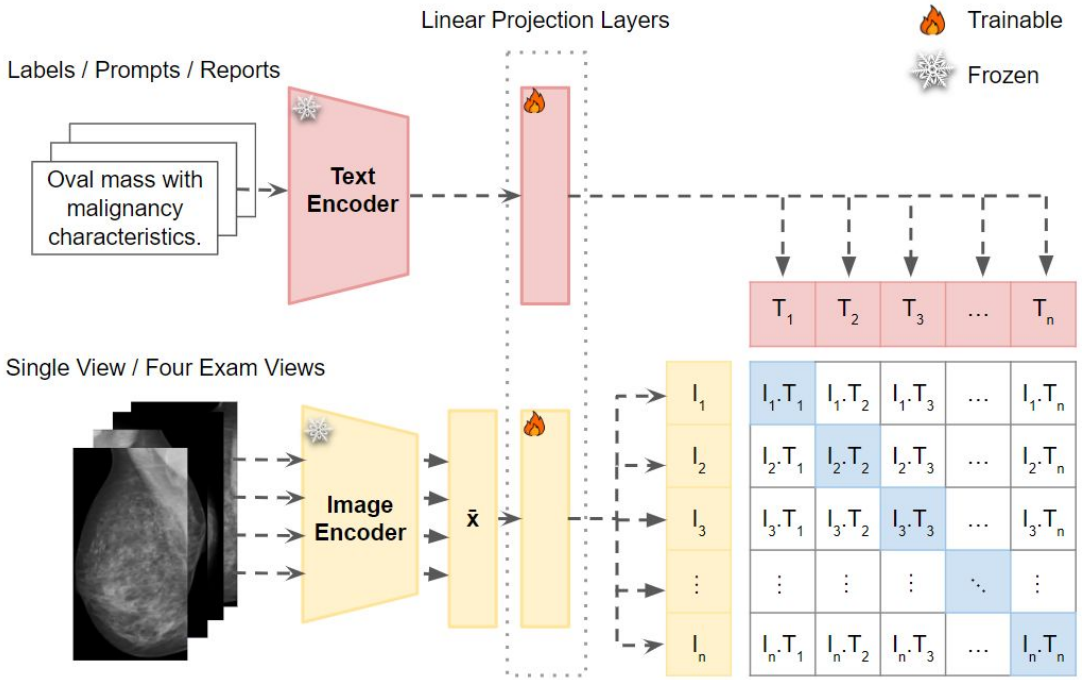
Can we allow the same network to accept image-prompt information? Thus, an image with multiple sentences.

↓

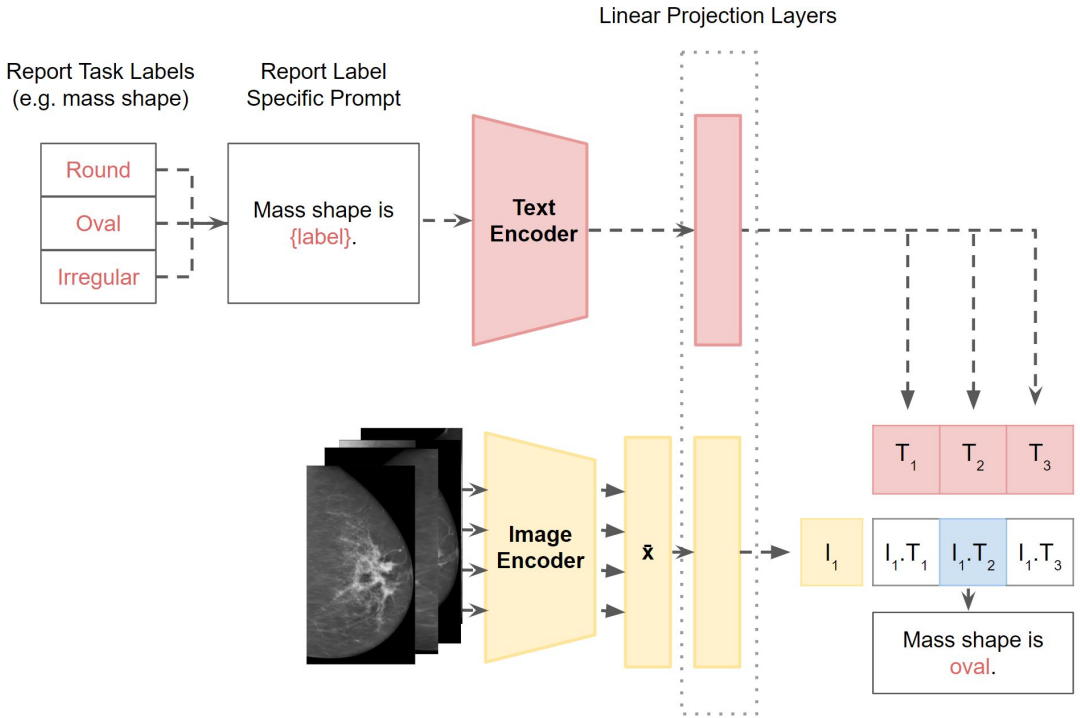
Can the network be trained on full exam-report information and understand the clinical meaning of the data?

Proposed Architecture: MMG-CLIP

(1) Training Architecture



(2) Report task zero-shot prediction



Experiments Results

Image Encoder Only

MMG-CLIP (Ours)

Experiments	Binary AUROC ↑				Average Multi AUROC (± std) ↑	
	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 (± 0.078)	0.8344 (± 0.089)

Table 1. Comparison of area under the ROC (AUROC) of classification tasks using one-vs-all classification evaluation on image level experiments.

- CNN image encoder (same encoder used in MMG-CLIP) was pre-trained on > 100k patient exams for malignancy task.
- Our network trained on image-label malignancy task outperformed a traditional CNN.

Experiments Results

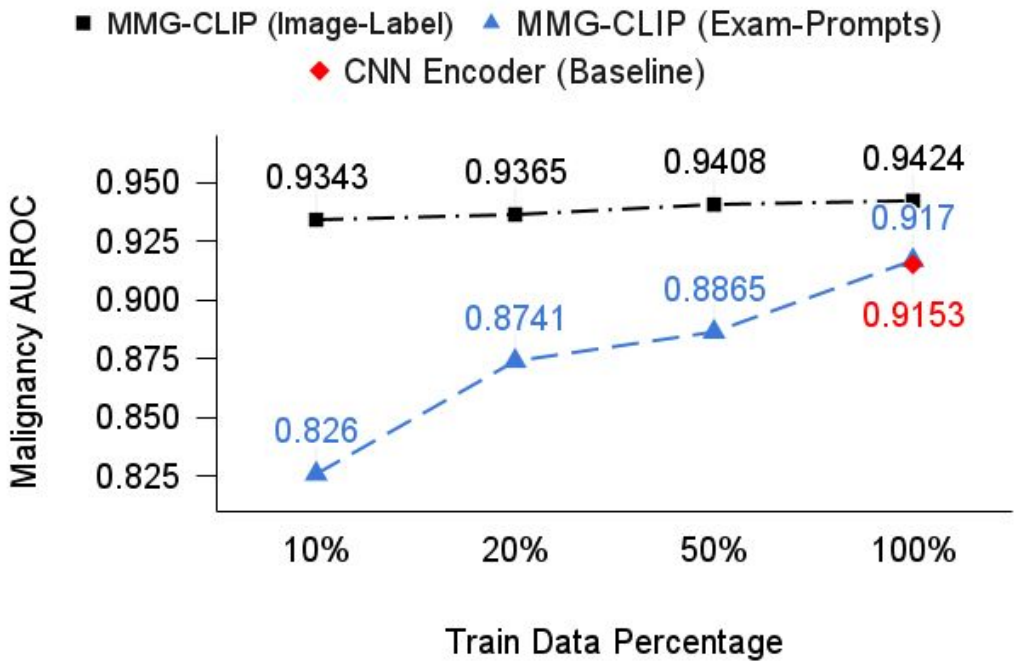
Table 2. Comparison of the average area under the ROC (AUROC) of different classification tasks using zero-shot classification evaluation on both image and exam level experiments.

	Average Binary Bootstrap Samples AUROC (95% CI) ↑				Average Multi AUROC (± std) ↑	
Experiments	Malignancy	Arch. Dist.	Mass	Calcification	Mass Shapes	Mass Margins
Image-Prompts	**0.931 (0.905-0.953)**	0.682 (0.554-0.808)	0.663 (0.564-0.755)	0.680 (0.639-0.719)	0.727 (± 0.120)	**0.715 (± 0.154)**
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 (± 0.079)	0.560 (± 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)	**0.848 (± 0.088)**	0.594 (± 0.094)
Exam-Prompts	0.916 (0.891-0.938)	**0.717 (0.620-0.804)**	**0.678 (0.603-0.743)**	**0.736 (0.701-0.772)**	0.700 (± 0.106)	0.639 (± 0.218)

- Adding additional visual and textual information (as full exam or reports) allows the network to perform on multiple tasks.
- Results obtained with our proposed prompts while training with images or exams outperformed training with reports.

Experiments Results

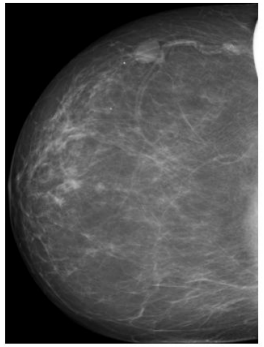
- Zero-shot performance, for either image-label or exam-prompts models on malignancy detection improves by training with more data.
- The models are effectively learning by additional data.

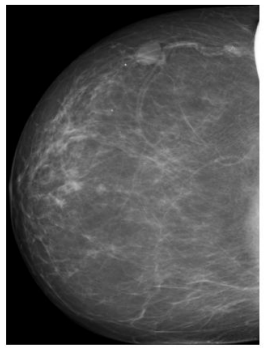


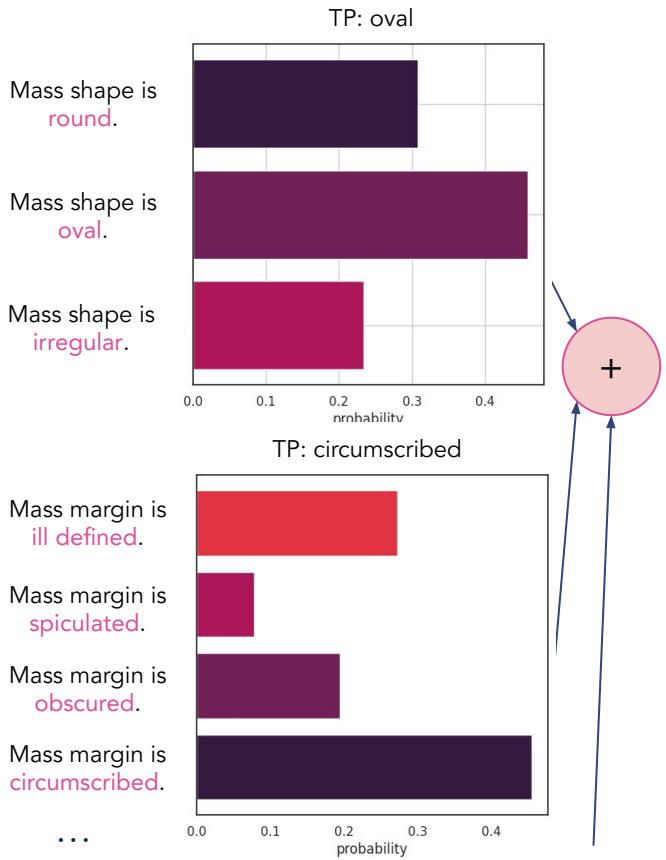
How is the report generation performed?

(1) BI-RADS Template (each color represents an inference task) → (2) Run several zero-shot predictions, for each task. → (3) Format and combine predictions to form a report.

MAMMOGRAPHY		
Breast composition	a. The breasts are almost entirely fatty b. There are scattered areas of fibroglandular density c. The breasts are heterogeneously dense, which may obscure small masses d. The breasts are extremely dense, which lowers the sensitivity of mammography	
Masses	Shape	Oval Round Irregular
	Margin	Circumscribed Obscured Microlobulated Indistinct Spiculated
	Density	High density Equal density Low density Fat-containing
Calcifications	Typically benign	Skin Vascular Coarse or "popcorn-like" Large rod-like Round Rim Dystrophic Milk of calcium Suture
	Suspicious morphology	Amorphous Coarse heterogeneous Fine pleomorphic Fine linear or fine-linear branching

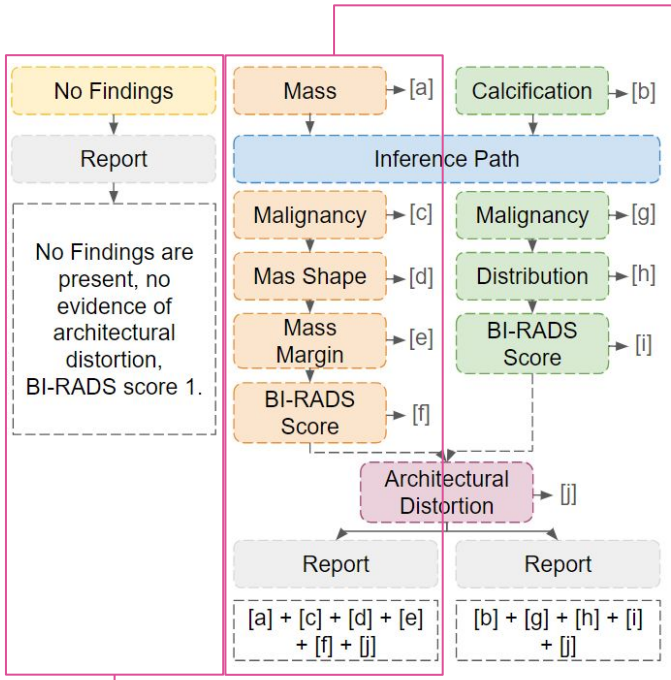


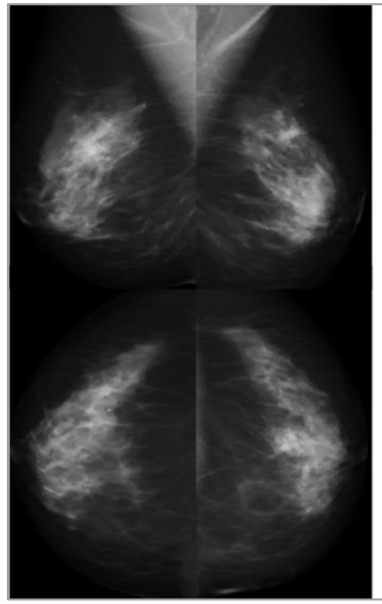




"Finding suggesting benign mass. Mass shape is oval. Mass margin is circumscribed. Mass density is equal. No evidence of architectural distortion."

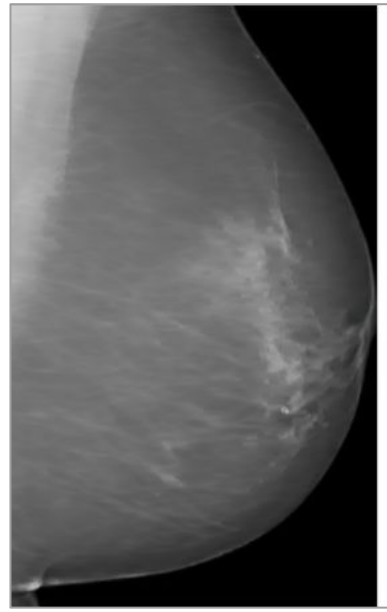
Generated Reports Examples





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."



Generated Report:

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

Green highlighted text represent correct predictions, yellow represents unknown labels, and red represents wrong predictions.

Contribution

- ➔ First work in the literature to utilize CLIP architecture for mammography report generation, using 4 exam views and multi-class prompts generation.
- ➔ Use-cases: image-to-text or text-to-image, flexible zero-shot predictions, report-generation, flexible binary/multi-class classifier.
- ➔ Limitations & future improvements:
  - ◆ Dataset -> Improve the textual information to have less unknowns, standardise reports.
  - ◆ Network -> Experiment pre-training the encoders on well structured prompts/reports.
  - ◆ Report generation -> Consider other decision approaches than taking the maximum probability.

Q&A

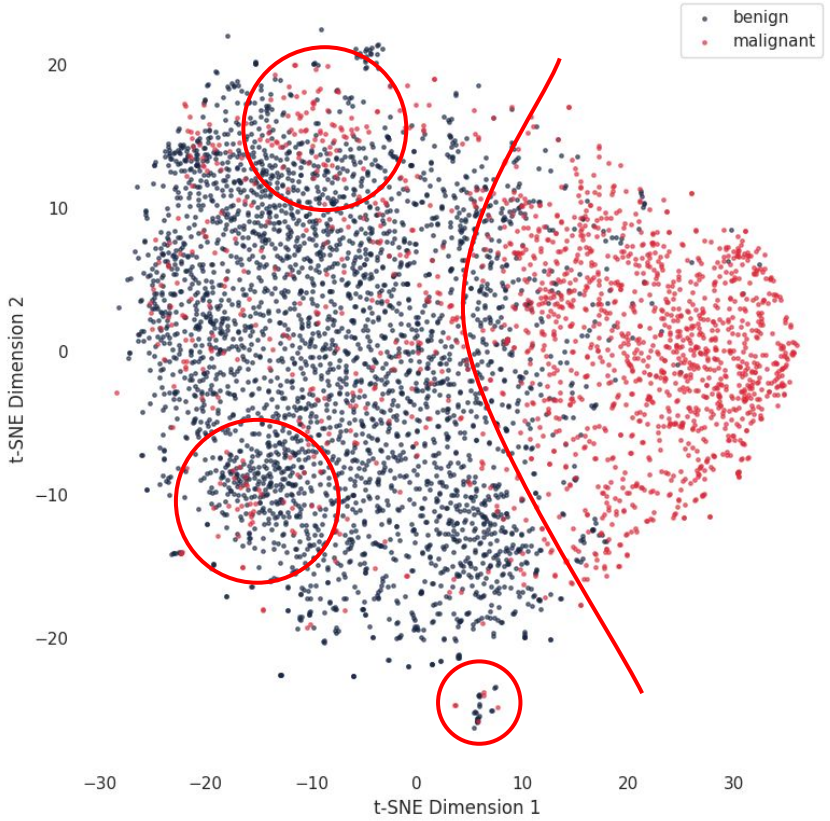
transpara®

By ScreenPoint Medical

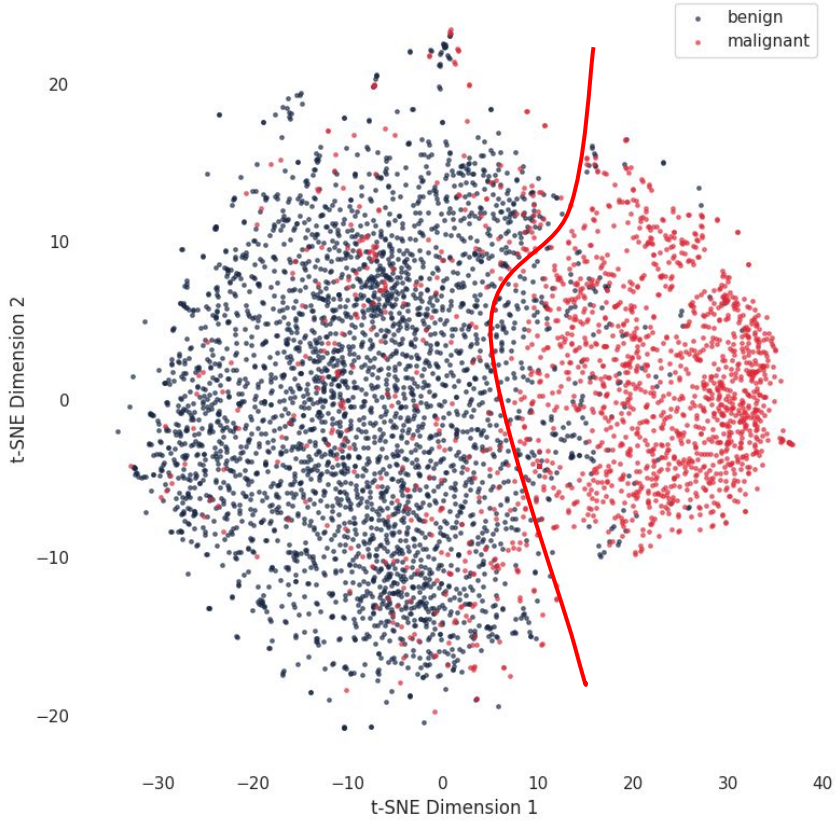
Experiments Results (t-SNE Analysis)

- Both CNN only (left) and CNN + Projection layers (ours, right) have good separation of malignancy features projected to a lower dimensional space.

- There is slight overlap using encoder only compared to adding projection layers.



(a) CNN Encoder (Baseline)

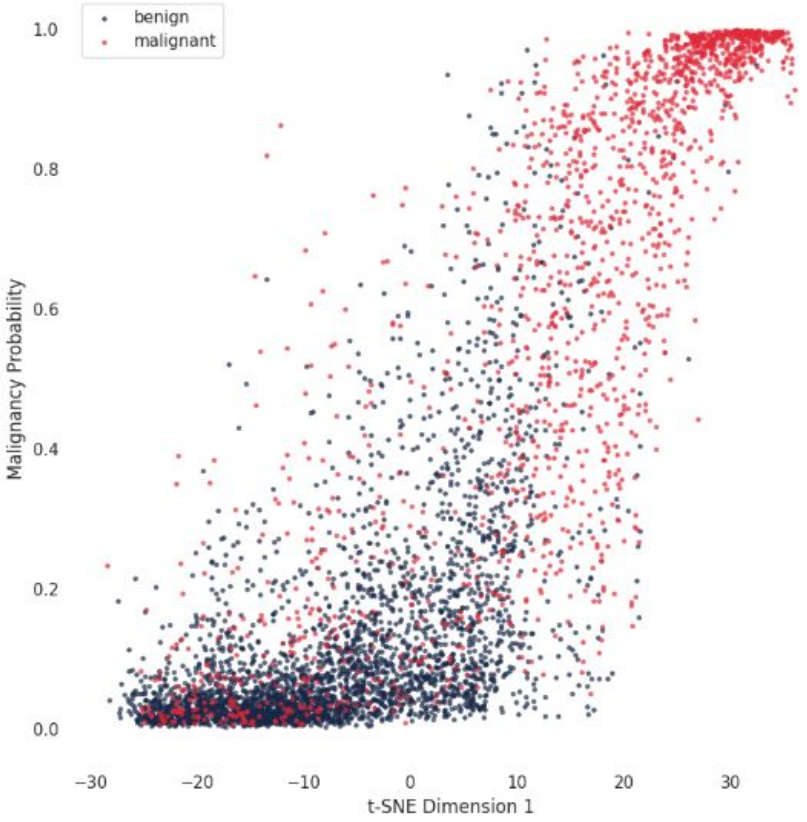


(b) CNN Encoder + Projection Layers

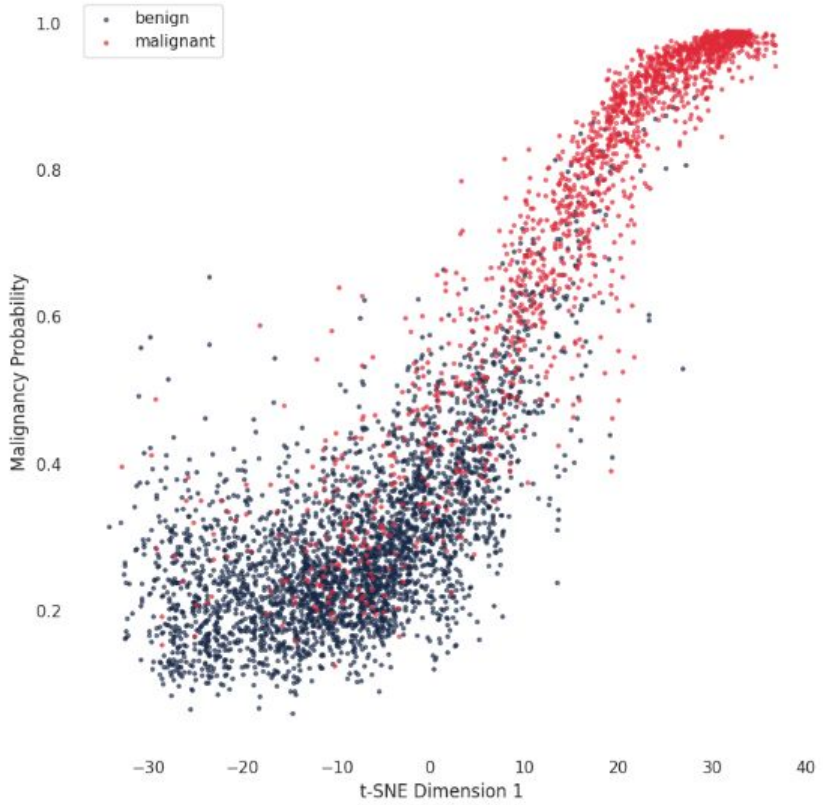
Experiments Results (t-SNE Analysis)

- Projecting the first dimension of t-SNE to the model malignancy probability indicates a positive correlation.

- Adding projection layers shows more distinct and reliable probability estimates for malignancy, and clearer separation.



(c) t-SNE dimension 1 vs malignancy probabilities for the CNN Encoder



(d) t-SNE dimension 1 vs malignancy probabilities for the CNN Encoder + Projection Layers

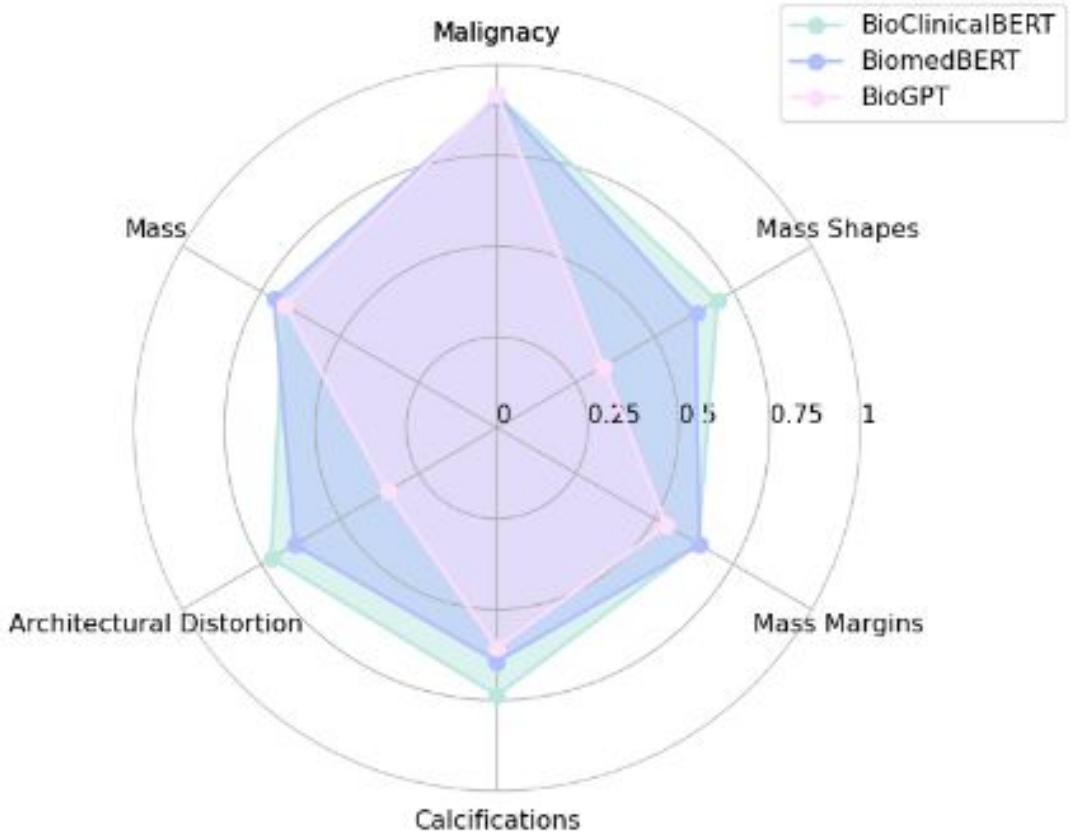
Ablation study on model architecture parameters

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

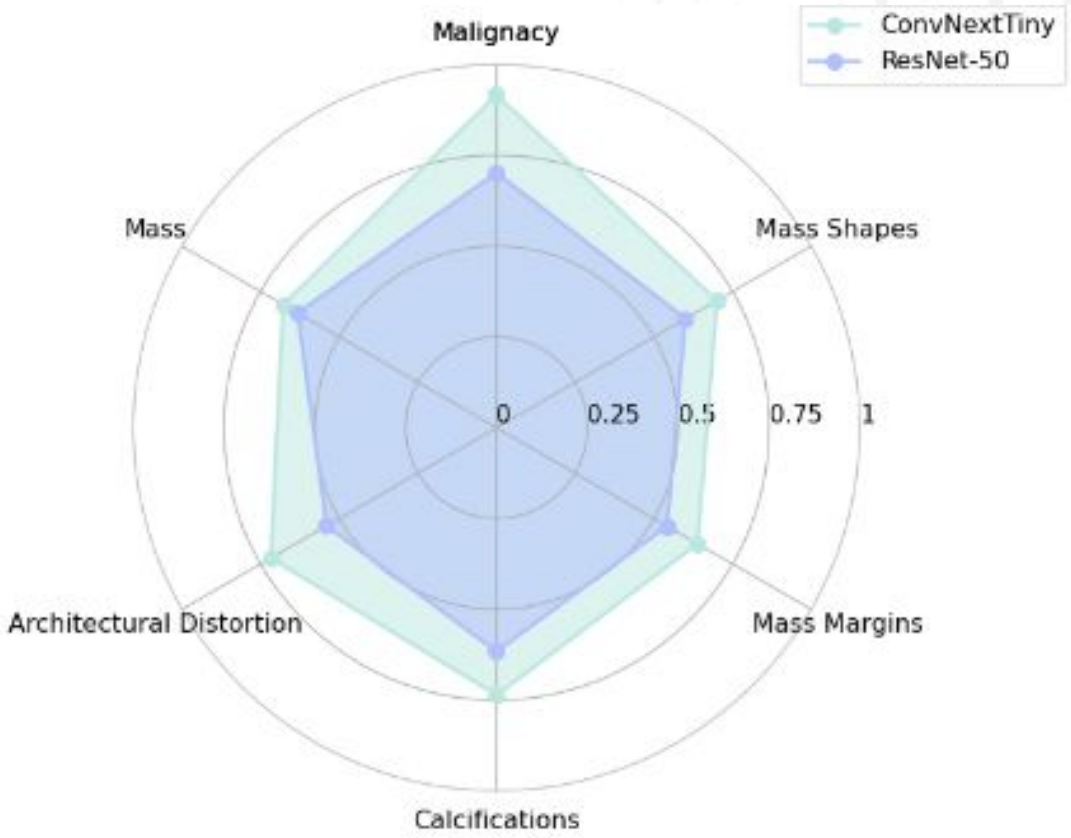
Experiments	AUROC (95% CI) ↑
MMG-CLIP	
w/ 1 proj. layers	0.893 (0.864-0.920)
w/ 2 proj. layers	**0.916 (0.891-0.938)**[a]
w/ 3 proj. layers	0.910 (0.882-0.933)
MMG-CLIP	
w/ batch size = 32	0.908 (0.883-0.933)
w/ batch size = 128	0.912 (0.885-0.936)
MMG-CLIP	
w/ seq. length = 384	0.910 (0.885-0.933)
w/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

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

Ablation study on pre-trained image and text encoders

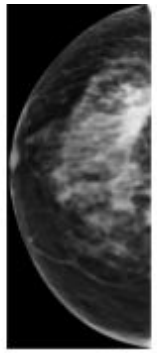


(a) Different LLM models performance as text encoders.

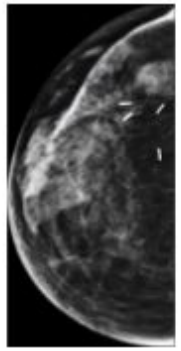


(b) Different vision models performance as image encoders.

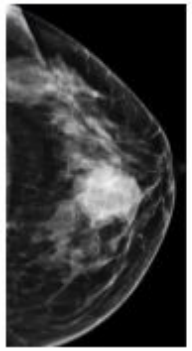
Data sampling for image-prompts experiment



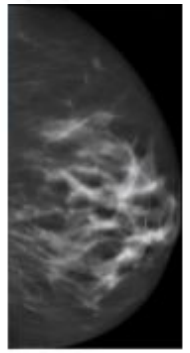
The mass displayed spiculated margins and irregular shape, suggestive of malignant features upon imaging.



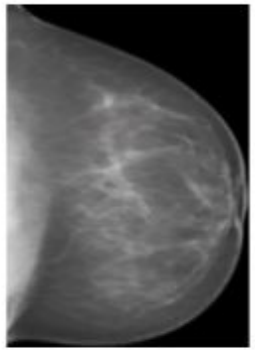
Suggestive of benign features upon imaging. Reported calcifications display benign characteristics



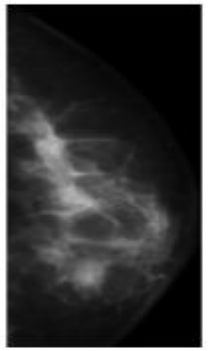
The mass demonstrated circumscribed margins and oval shape, indicating a likely benign etiology.



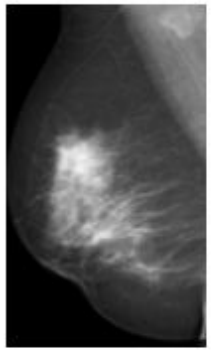
The mass exhibited ill defined margins and irregular shape, suggesting potential malignant pathology.



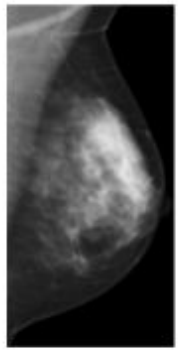
Suggestive of a benign lesion.



The present mass appeared obscured margins and round shape, indicating potential malignant characteristics.

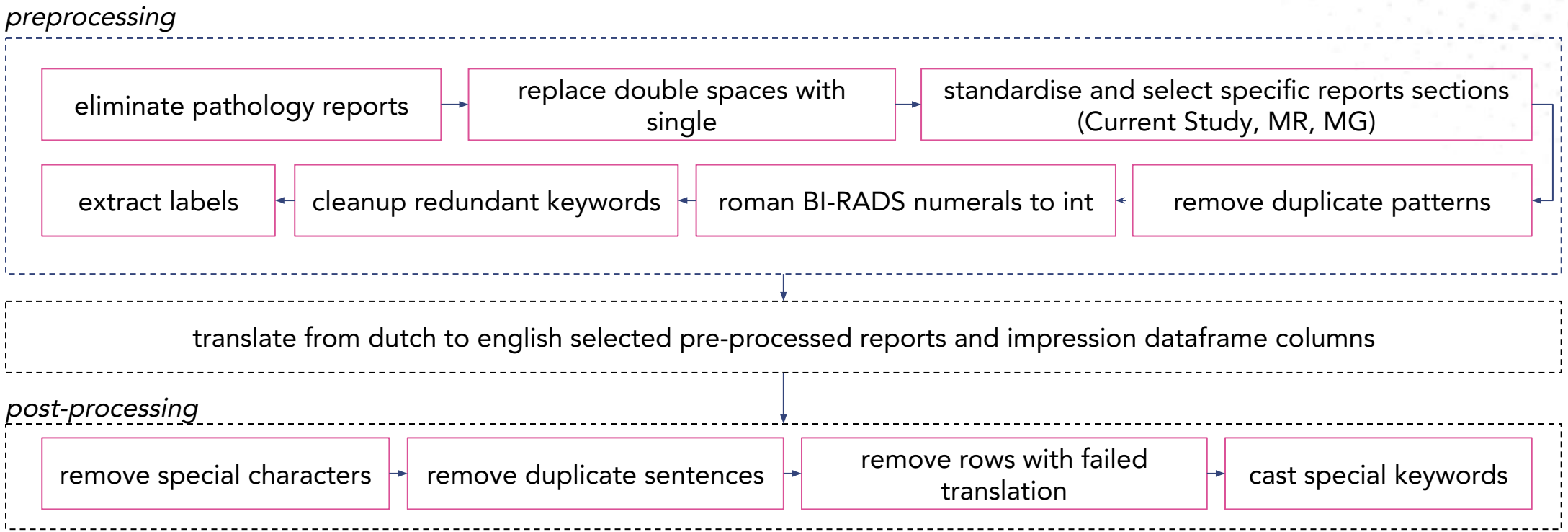


The of the mass seen on imaging were ill defined margins and irregular shape, prompting concern for malignant.



Indicative of potential benign. Identified calcifications exhibit features indicative of benign.

Text Reports Preprocessing



Interpreting model outputs

Model output logits (un-normalized probabilities)

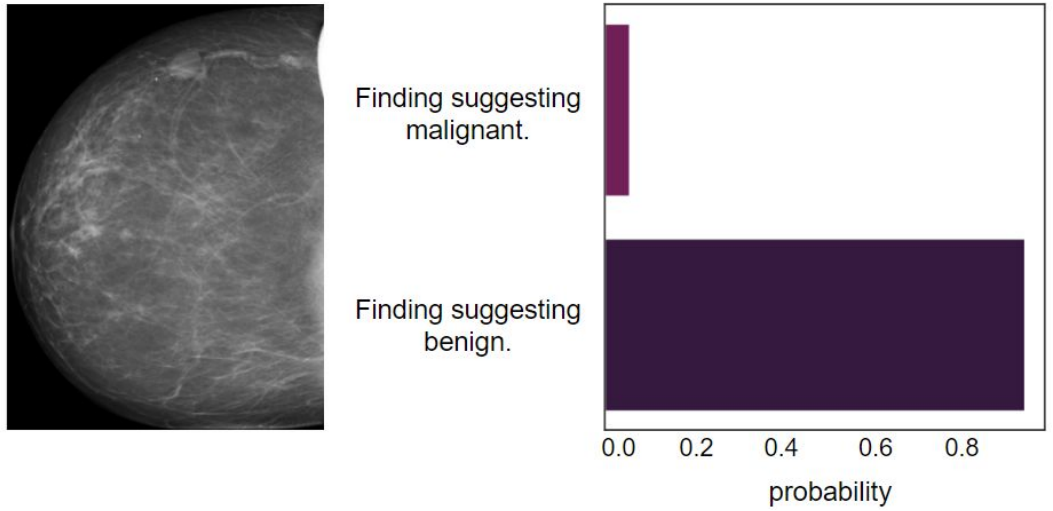
$I_1.T_1$ = -0.0331	$I_1.T_2$= 2.912
Finding suggesting malignant.	Finding suggesting benign.

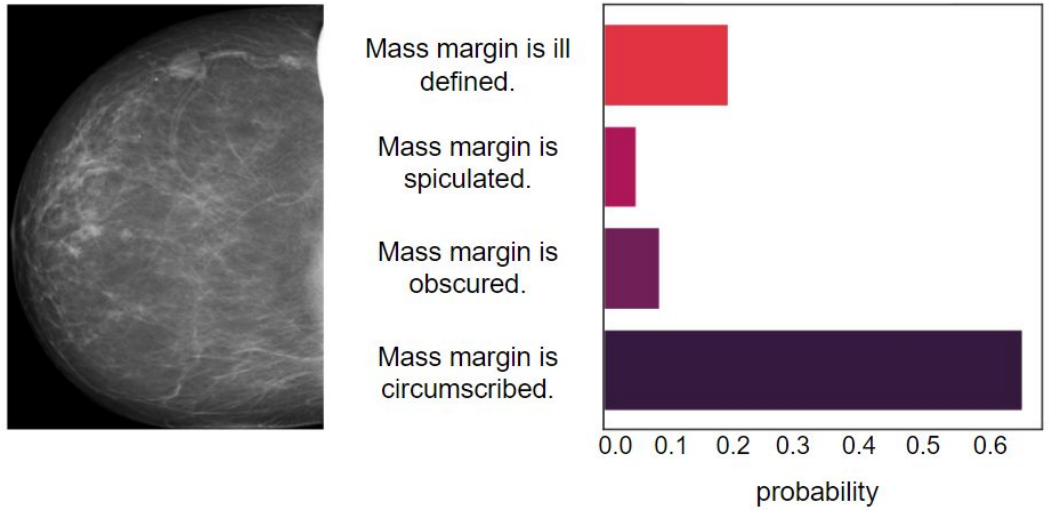
→ *softmax layer* →

$$\sigma(\vec{z})_i$$

Normalized probabilities

$I_1.T_1$ = 0.0499	$I_1.T_2$= 0.9500
Finding suggesting malignant.	Finding suggesting benign.





Dataset description

Labels and BI-RADS tasks

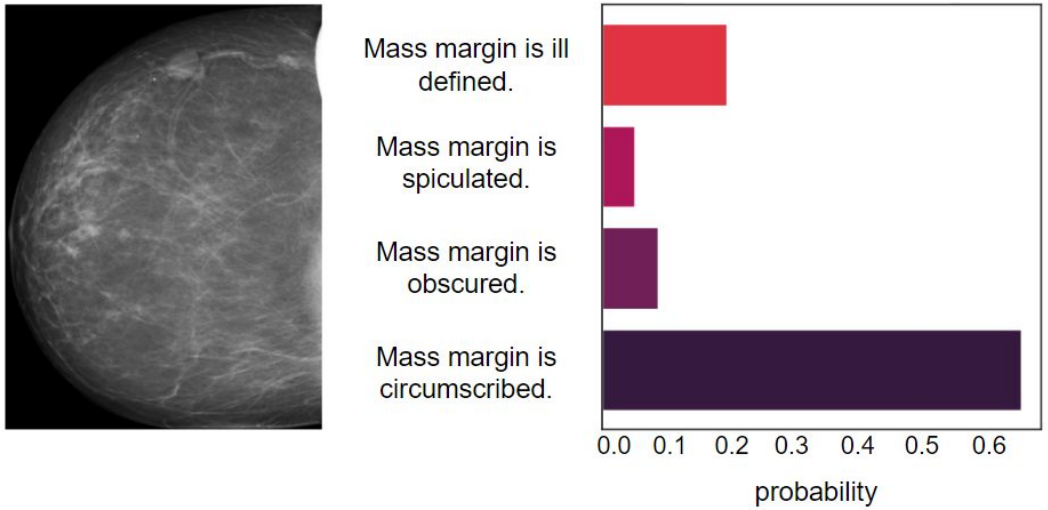
Label Group	Labels Names	Count
Malignancy	Benign	3311
	Malignant	1653
Mass Margins	Unknown	2467
	Ill defined	1095
	Obscured	697
	Spiculated	484
	Circumscribed	221
Mass Shapes	Unknown	2466
	Irregular	1218
	Round	681
	Oval	599
Architectural Distortion	Normal	4842
	Distortion	122
Calcification	No Calcification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

Datasets Split

Dataset	Split	Count
Image-Label *or*	Train	3474
Image-Prompts	Valid	1490
	Test	745
Exam-Reports *or*	Train	1282
Exam-Reports + Prompts *or*	Valid	550
Exam-Prompts	Test	745

Zero-shot evaluation input prompts

Out label specific zero-shot input evaluation prompts.



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.

Experiments Parameters

Parameter	Value
Embedding dimension	512
Early stopping patience	5
Tokenizer sequence length	256
Warm-up epochs	0.1
Trainable epochs	30
Initial Learning rate	5e-5
Weight decay	1e-4

Components	Name
Scheduler	Cosine-annealing learning rate
Optimizer	AdamW
Text pooling	[EOS] global representation

Experiment	# Trainable Linear Layers	Batch Size
Image-Label (binary/multi)	1	32
Image-Prompts	1	64
Exam-Reports	2	64
Exam-Prompts	2	64
Exam-Reports+Prompts	2	64

Comparison to LLMs

1. *Output can be very specific and controlled*

   Unlike LLMs that can deviate from specific requirements and generate un-related text; using our approach for a specific guidelines (BI-RADs) can generate more specific and accurate text.

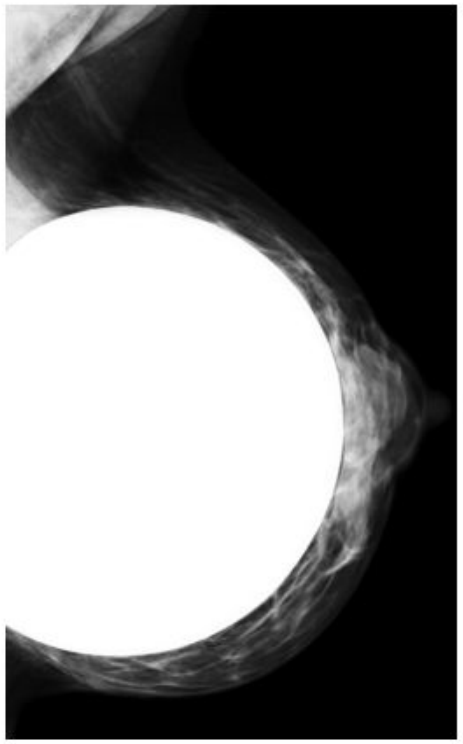
2. *Flexible and has more potential applications*

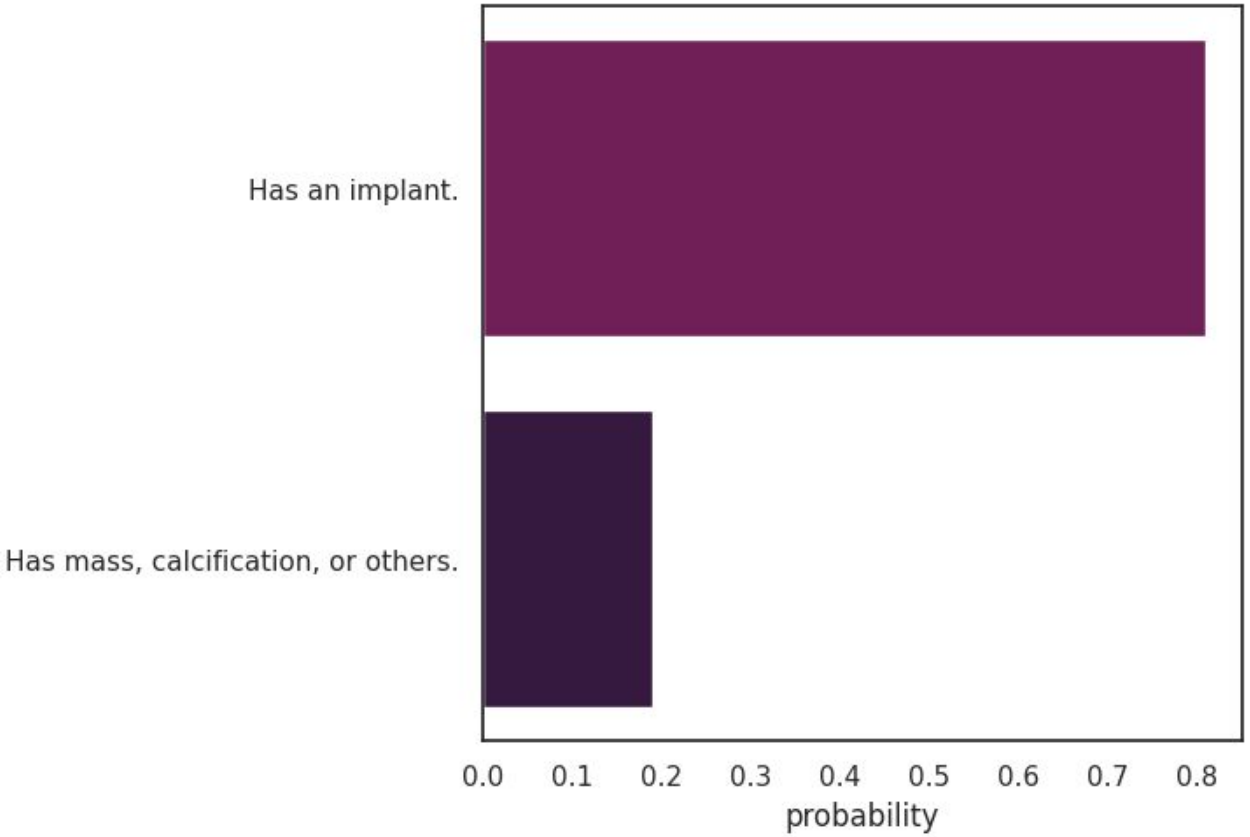
   LLMs can generate text, but can't be used for zero-shot classification when there is lack in data. Another advantage is that MMG-CLIP can be used as a general zero-shot classifier for binary/multi-class tasks even when it was not trained on specific task (see implant classifier example in the appendix).

3. *Can be tailored to any small scale image-text or image-label dataset*

   Unlike many LLMs, MMG-CLIP can be trained on small sample size datasets.

The network was not trained on any information related to implants; but was able to detect in several examples.

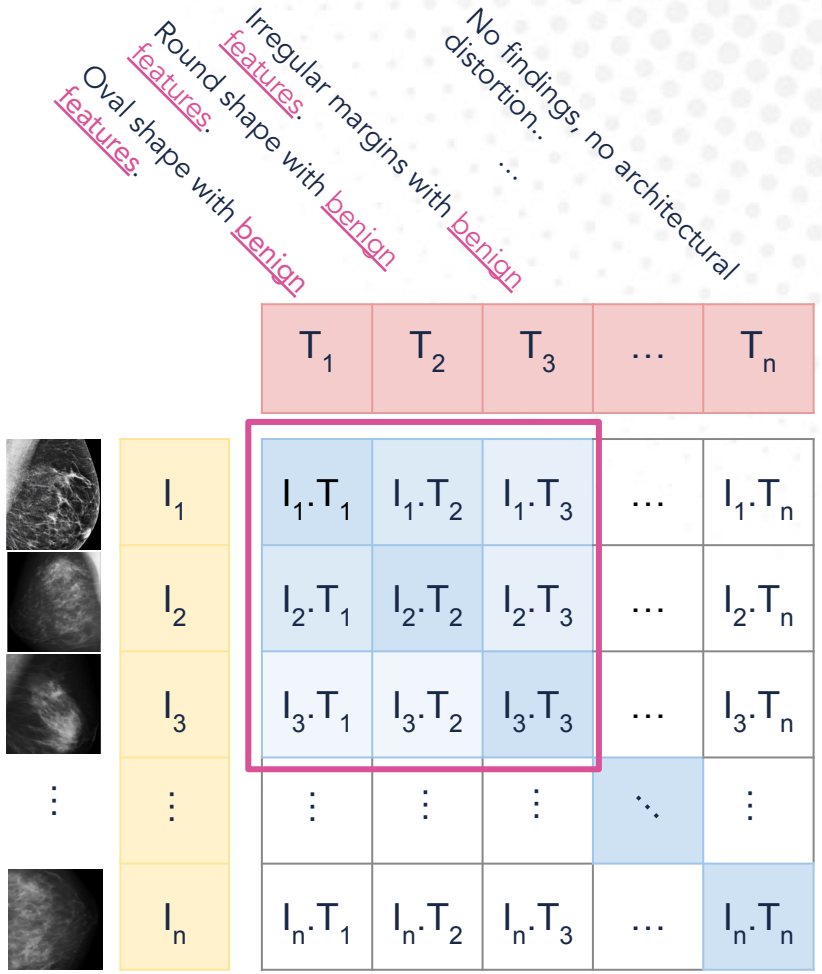




Limitations and future work (1/3)

1. Embedding Matching

- For small-scale medical images, where the text can contain similar information for multiple samples, CLIP InfoLoss might not be the best choice as the network will match not only the TP pair, but also the FN pairs.
- Medical images can be paired to multiple reports sections, and this would result to high loss during training.
- Future Work: Use attention mechanism, new loss, region-wise matching instead of global.

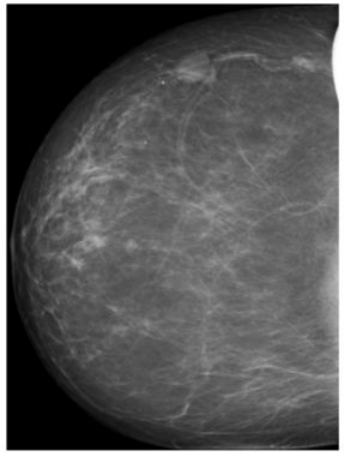


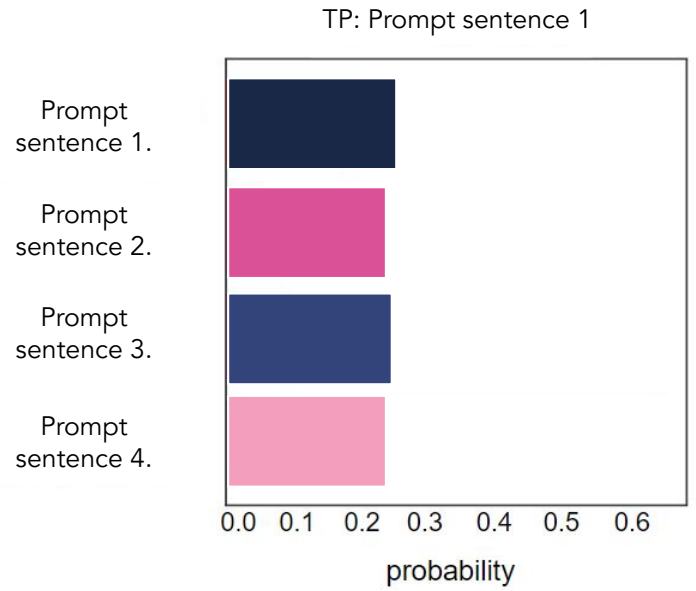
Limitations and future work (2/3)

2. Report Generation

- The model is not very confident with the result as it outputs similarities that are very close to each other. Taking the maximum similarity as the correct output might not be the best approach.

- Future Work: consider other decision making approaches like thresholding or reject outputs with similar probabilities.





Limitations and future work (3/3)

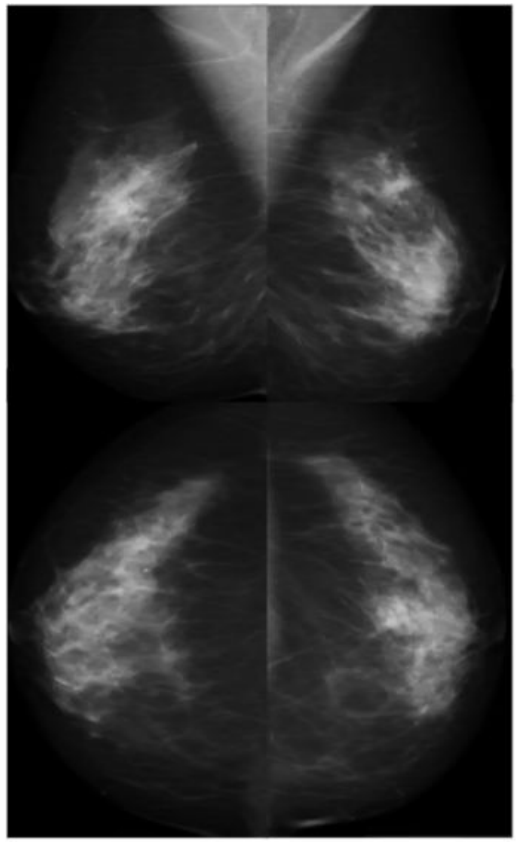
3. Pre-training

- The network performance heavily relies on the pre-trained encoders. Having domain-specific pre-trained encoders will significantly improve the results.

4. Model Architecture

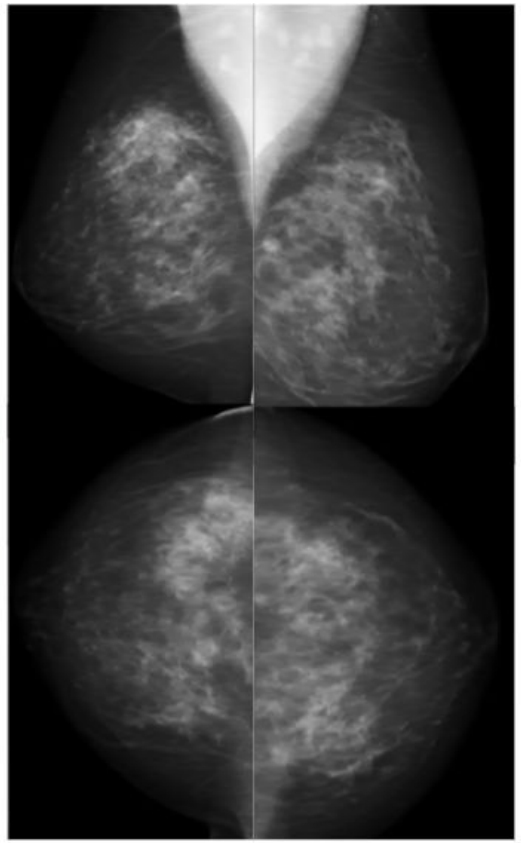
- Averaging the exam images features loses information related to the location of the region of interest (e.g. mass in left MLO view).
- Future Work: Experiment other concatenation approaches.

Generated Reports Examples



Generated Report:

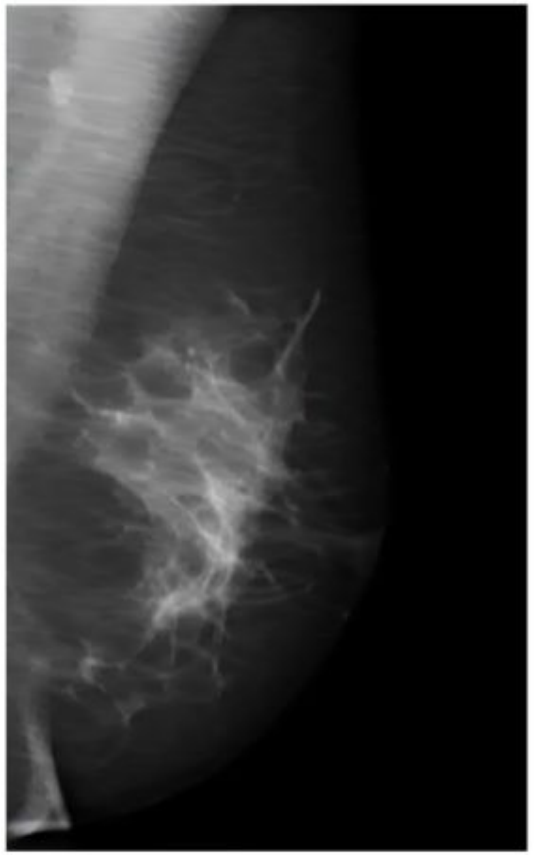
"The mass demonstrated spiculated margins and irregular shape, prompting further evaluation for malignant features, this concludes assigning a BIRADS score of 0. No evidence of architectural distortion was noted on mammography."



Generated Report:

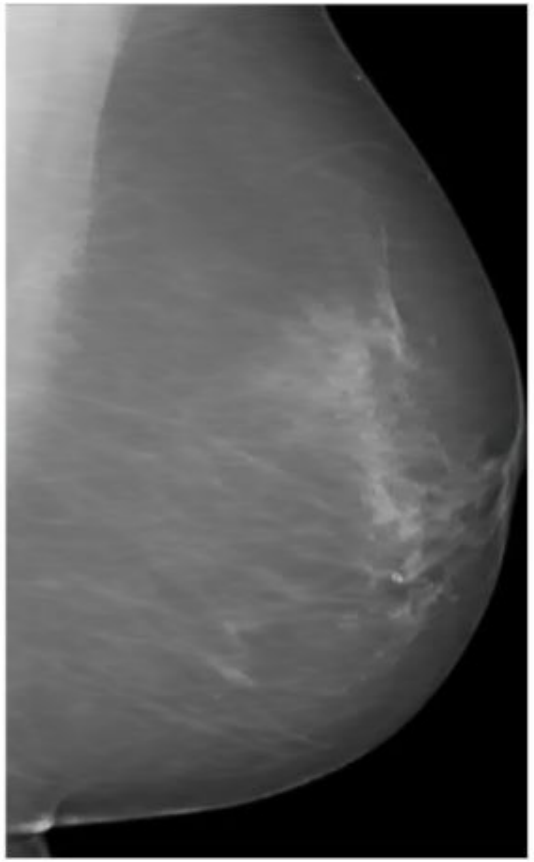
"Observed calcifications appear benign with regional distribution, assigned BIRADS score 3 for clinical management. The presence of architectural distortion on mammography necessitated careful evaluation."

Generated Reports Examples



Generated Report:

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



Generated Report:

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