MMBCD: Multimodal Breast Cancer Detection from Mammograms with Clinical History

Kshitiz Jain[1], Aditya Bansal[2⋆], Krithika Rangarajan[1,3], and Chetan Arora[1⋆⋆]

[1] Indian Institute of Technology Delhi, Delhi, India
[2] Thapar Institute of Engineering and Technology, Punjab, India
[3] All India Institute of Medical Sciences Delhi, India

Abstract. Mammography serves as a vital tool for breast cancer detection, with screening and diagnostic modalities catering to distinct patient populations. However, in resource-constrained settings, screening mammography may not be feasible, necessitating reliance on diagnostic approaches. Recent advances in deep learning have shown promise in automated malignancy prediction, yet existing methodologies often overlook crucial clinical context inherent in diagnostic mammography. In this study, we propose a novel approach to integrate mammograms and clinical history to enhance breast cancer detection accuracy. To achieve our objective, we leverage recent advances in foundational models, where we use `ViT` for mammograms, and `RoBERTa` for encoding text based clinical history. Since, current implementations of `ViT` can not handle large $4K \times 4K$ mammography scans, we device a novel framework to first detect region-of-interests, and then classify using multi-instance-learning strategy, while allowing text embedding from clinical history to attend to the visual regions of interest from the mammograms. Extensive experimentation demonstrates that our model, `MMBCD`, successfully incorporates contextual information while preserving image resolution and context, leading to superior results over existing methods, and showcasing its potential to significantly improve breast cancer screening practices. We report an (Accuracy, `F1`) of (0.96,0.82), and (0.95,0.68) on our two in-house test datasets by `MMBCD`, against (0.91,0.41), and (0.87,0.39) by `LLaVA`, and (0.84,0.50), and (0.91,0.27) by `CLIP-ViT`; both state-of-the-art multi-modal foundational models.

Keywords: Mammography, Clinical History, Screening, Diagnostic

1 Introduction

Screening vs. diagnostic mammography. We observe that most breast cancer detection techniques are based on screening mammography, and often overlook the crucial inputs of mammogram findings and patient history. This omission of clinical history can undermine the reliability and performance of deep neural network models, as they fail to incorporate essential diagnostic features

⋆ Work done as an intern at Indian Institute of Technology Delhi
⋆⋆ Corresponding author

valued by trained radiologists. For the readers of this paper who may not have a medical background, screening mammograms are similar to annual preventive exams based on the risk factors, and may not involve any specific complaint from a patient. On the other hand, a diagnostic mammogram is asked by a doctor based on any abnormal indications or signs of breast cancer symptoms. A diagnostic mammogram often contains detailed patient history in text form.

Clinical history in breast cancer detection. Clinical history of a patient holds immense value for radiologists, as it provides crucial context for interpreting imaging studies. Without pertinent clinical information, the significance of an imaging study may be diminished [2]. Collecting patient history is straightforward and can be completed by patients themselves, either independently or with minimal assistance. For mammograms, history is often gathered through a cost-effective questionnaire that elicits details about breast-related symptoms (such as a lump, discharge, pain, etc), prior breast cancer diagnoses, surgeries, radiation therapy, and other risk factors such as the family history of breast cancer, concurrent illnesses like as other cancers. Integrating patient history into neural network algorithms holds promise for enhancing detection accuracy.

Related works. Kooi *et al.* [11] investigated the potential impact of patient age on model performance in their study, but did not find significant contribution from patient's age towards their model's performance. Tang *et al.* [19] explored the integration of clinical history in a weakly supervised framework to enhance breast cancer detection in mammograms. However, their methodology relied on costly annotations of segmentation masks. Tang *et al.* did not release their dataset or share code or trained models, precluding direct comparison with our methodology. Zheng *et al.* [22] utilized tabular data alongside CT images to predict patient survival outcomes, while Liu *et al.* [13] leveraged the CLIP foundation model for organ segmentation in CT scans. Hager *et al.* [6] explored aligning cardiac MR images with patients' routine clinical data. Wang *et al.* [20] introduced MedCLIP, a foundation model trained on extensive image data, demonstrating the capability of zero-shot predictions. Recent advancements in vision literature have witnessed notable enhancements in the performance of models such as Vision Transformers (`ViT`) [5] and `ResNet` [7]. We harness these models, pre-trained on vast image datasets, for their learned representations.

Challenges and our approach. We observe that foundational models, primarily designed for natural images, encounter challenges when applied directly to breast cancer detection. Mammograms, with a resolution of $4K \times 4K$, often depict minute cancerous lesions spanning just a few hundred pixels. However, current deep neural network models necessitate resizing the images to a smaller dimension, leading to diminished performance, particularly in detecting small cancers [16]. Our method addresses the challenge by treating a mammogram, not as a single sample, but as a set of Regions of Interest (`ROIs`), where cancer may be present, and reformulate the problem as a multi-instance learning [4,9].

Contributions. (1) Changing the attention of the community from screening to diagnostic mammography, we introduce multi-modal breast cancer detection model, leveraging rich information in the clinical history. **(2)** We leverage recent

advancements in foundational models, and propose a pluggable architecture, which fuses visual embeddings from ViT, and textual embeddings from RoBERTa models for accurate cancer detection. In the process we overcome several challenges around handling of large mammograms by foundational models, by using a novel region-of-interest based architecture along with multi-instance-learning. The proposed architecture not only allows to use full resolution mammograms for training, and inference, but also helps our model to focus on salient regions (extremely important to learn from small datasets). **(3)** Our MMBCD model successfully incorporates contextual information while preserving image resolution and context, leading to superior results over existing methods as already mentioned in the abstract of the manuscript.

2 Dataset

Our dataset comprises two distinct AIIMS datasets collected from different departments within our institution. The first dataset encompasses diagnostic mammography screenings, where patients presenting with breast-related concerns undergo mammography examinations. In contrast, the second dataset consists of opportunistic mammography screenings, involving patients with other cancer types, individuals with a history of breast cancer, and those seeking general screening services. Both datasets were acquired using similar imaging machines, and clinicians adhered to a standardized protocol, namely the ACR BIRADS[1] lexicon, when documenting indications and comments during patient diagnosis. Due to variations in patient demographics between the two datasets, the nature of comments provided by clinicians also exhibits differences. The first dataset predominantly includes complaints from patients, whereas the second dataset contains information concerning the patients' medical history.

AIIMS 1. This dataset comprises 3,816 studies, each containing digital mammograms of patients along with their corresponding clinical background data in textual format. The dataset encompasses three types of studies: Bilateral Mammography, Left Mammography, and Right Mammography, with bilateral mammography being the predominant type. We partitioned this dataset into training and testing sets, consisting of 2,965 (571 malignancies) and 851 (148 malignancies) studies, respectively. The split was based on acquisition dates, with the training set containing studies from May 2013 to March 2016, and the testing set comprising consecutive studies from January 2018 to March 2019. This split was chosen to evaluate our model's performance under real-world testing scenarios, enhancing its reliability.

AIIMS 2. This dataset consists of 583 (58 malignancies) studies obtained from a different department within our hospital. Mimicking the specifications and procedures of AIIMS 1 dataset, this dataset serves as an external validation set to demonstrate the generalization capability of our model. This dataset is acquired from the studies that happened between January 2020 to December 2020. We are planning to release this dataset in near future, for future research and optimal comparison.

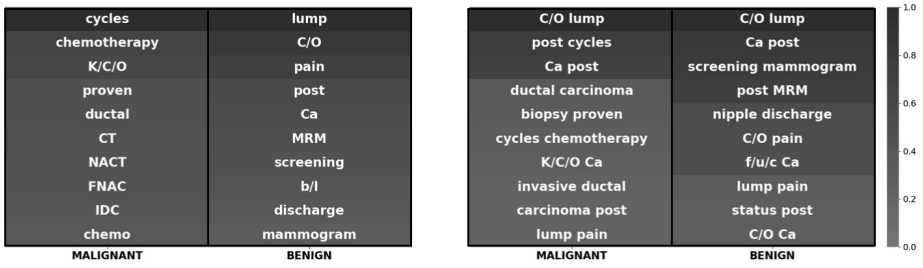

Fig. 1. Analysis of clinical history in our dataset. (a) Words with high `tf-idf` scores unique to each diagnosis. (b) Bigrams with the highest frequencies. Scores for the mentioned keywords are normalized for heatmap visualization.

Clinical History. The clinical history included in both AIIMS datasets consists of textual notes and patient complaints recorded by medical professionals. Each clinical history typically ranges from 1 to 45 words, with an average length of 8 words. These notes reflect the doctor's observations and speculations regarding the patient's complaints, excluding any pathology reports. The textual data encompasses descriptions of prior breast cancer concerns and patient discomforts. For patients undergoing regular screening, the clinical history often contains the term "screening". To process this data, we utilize the built-in tokenizer provided by `RoBERTa` [14]. In Fig. 1, we highlight class-specific important unigrams and bigrams occurring in our training split.

3 Methodology

In this section, we outline various components of our model, consisting of three distinct parts: **(1)** We introduce detection module to generate Region of Interest (`ROI`) candidates crucial for our pipeline. **(2)** We generate visual embeddings from the `ROIs`, and textual embeddings from the clinical history. **(3)** We develop a cross-attention module to fuse visual and textual embeddings.

3.1 ROI extraction

Large size of mammograms, $4K \times 4K$, poses a significant challenge, even for large vision foundational models (`LVFMs`), to effectively capture relevant features. Given that cancerous lesions in mammograms are often small, and the majority of the image area is benign, and looks similar across samples, it essentially converts our problem to a fine-grained visual classification task, where focusing on salient regions becomes extremely important. To help a foundational model in our architecture to focus on salient regions, we employ an object detection modules[10,17,21] trained to delineate bounding boxes around cancerous regions within mammograms. However, training such models typically necessitates costly bounding box annotations. To overcome this limitation, we curate

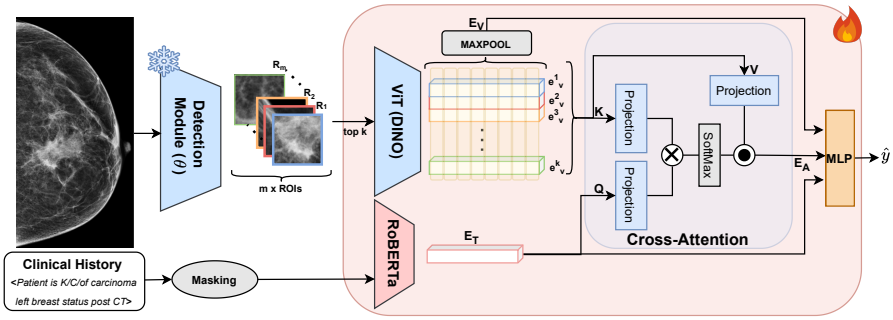

Fig. 2. Proposed architecture. Our framework utilizes a cross-attention layer to attend to top K `ROI` embeddings by the textual representation of the clinical history. Our findings reveal the synergistic impact of textual, visual, and cross-attention embeddings on the accuracy of breast cancer detection.

a small subset comprising approximately 390 malignant mammograms from our training dataset. Additionally, we utilize 540 benign images, which do not require bounding box annotations. This approach enables us to effectively train our module to provide `ROI` annotations over mammograms. To perform `ROI` extraction, we utilize `FocalNet-DINO` [21] (θ) as our detection module. Given a set of input images $I^n_{i=1}$, the module outputs m bounding box predictions for each image I_i, denoted as: $\{R^m_{j=1}\}^n_{i=1} = \theta(I^n_{i=1})$. Here, θ represents the parameters of the `FocalNet-DINO` model. Each $R^m_{j=1}$ prediction corresponds to a bounding box around a potential cancerous region within the respective image I_i.

3.2 Combining Text and Image: MIL with Late Fusion

Vision module with multi-instance learning. To train our `ViT` based vision encoder, we treat the `ROIs` extracted from mammograms as a set $\{R^m_{j=1}\}_i$. For malignant images, this set comprises both malignant and benign regions of the mammogram, whereas for benign images, it consists solely of benign regions. In this setup, we leverage classical multi-instance learning (`MIL`) strategy. We train our image encoder (V) by inputting all the `ROIs` to it and extracting visual embeddings ($\{e_v\}^m_{j=1}$) from each `ROI`. Subsequently, we aggregate these embeddings to create a collective representation (E_V) that encapsulates contextual information from across the entire mammogram. This aggregation is achieved using Max-Pooling over all the output embeddings from the vision encoder:

$$\{e_v\}^m_{j=1} = V(R^m_{j=1}); \qquad E_V = \text{Max-Pool}(e^m_{j=1}, dim = 1). \tag{1}$$

The rationale behind using Max-Pooling is to ensure that we capture all relevant features present in any of the `ROIs`. In contrast, taking the mean of the final embeddings may dilute the embedding if the number of benign embeddings outweighs the number of malignant embeddings, potentially diminishing the representation of malignant features. Specifically, we utilize the Vision Transformer (`ViT`) architecture from the open-source `DINO` [3] repository.

Text encoder. For the text encoder, we utilize a sequence classification module to encode the clinical history ($H^n_{i=1}$) of the patients. In our experiments, we employ another text foundational model, `RoBERTa` [14] to extract sentence embeddings. We denote `RoBERTa` as (T), and use output corresponding to `[CLS]` token as sentence embeddings (E_T). During experimentation, we observed that certain keywords (as depicted in Fig. 1) appear exclusively in specific classes, potentially leading to overfitting of our model to these keywords. To mitigate this issue, we identify 100 most relevant keywords in our corpus using `tf-idf` scores [18] and randomly mask these keywords during training:

$$\{E_T\}^n_{i=1} \;=\; T(H^n_{i=1}). \tag{2}$$

It is important to note that our model operates under a single-view classification paradigm. For patients diagnosed with malignancy, during training, we do not include clinical history from the unaffected breast.

3.3 Cross-attention between clinical history and image `ROIs`

To integrate both the visual ($\{e_v\}^m_{j=1}$) and clinical history (E_T) embeddings, we train a cross-attention module between these embeddings. In this setup, the clinical history E_T serves as the query, while each $\{e_v\}_j$ acts as the key and value. Consequently, this module produces a single unified embedding (E_A), where each `ROI` is attended to by the patient's clinical history. The attention mechanism can be described as follows:

$$E_A = \text{Attention}(E_T, \{e_v\}^m_{j=1}) = \text{softmax}\left(\frac{E_T \cdot (\{e_v\}^m_{j=1})^T}{\sqrt{d_k}}\right)\{e_v\}^m_{j=1} \tag{3}$$

Here, d_k represents the embedding size. This embedding plays a crucial role in the model, as it addresses the challenge of detecting small cancerous lesions present in only a few of the `ROIs` obtained from the detection module. By learning the relationship between a patient's history and the `ROIs`, the model can determine which `ROIs` best represent the patient's clinical background. Finally, we concatenate the obtained embeddings (E_V, E_A, *and* E_T), and pass them through a Multi-Layer Perceptron (`MLP`) to generate the malignancy score ($\hat{y}$).

4 Experiments and Results

Implementation Details. Our detection model is initialized with `COCO` weights, while vision encoders leverage various publicly available pre-trained weights. We employ Non-Maximum Suppression (`NMS`) with an Intersection over Union (`IoU`) threshold of 0.1 on detection model's output. This facilitates obtaining non-overlapping bounding boxes, enhancing contextual information for subsequent processing stages. We use Cross-Entropy loss to back-propagate the classification loss. In training the text encoder, masking ratio is set to 20%, learning rate is set to 1e-06 for whole network, and trained for 100 epochs. This training process

Table 1. Comparison of multi-modal approaches for breast cancer detection, including various CLIP models and LLava-1.5 using the LORA method.

Model	Weights	AIIMS 1			AIIMS 2		
		Acc.	F1	AUC	Acc.	F1	AUC
CLIP-R50[15]	ImageNet	0.69	0.36	0.820	0.64	0.22	0.769
CLIP-R50[15]	CLIP	0.76	0.34	0.759	0.73	0.21	0.694
CLIP-ViT[15]	ImageNet	0.83	0.47	0.856	0.69	0.20	0.699
CLIP-ViT[15]	DINO	0.84	0.5	0.895	0.91	0.27	0.741
LLaVA[12]	LCS-558K	0.91	0.41	-	0.87	0.39	-
OURS		**0.96**	**0.82**	**0.973**	**0.95**	**0.68**	**0.950**

is executed on a server equipped with 8 NVidia A100 GPUs, each possessing 80GB of memory. To ensure complete reproducibility, we release the source code of our model. [4]

4.1 Comparison with multi-modal foundational models

We evaluate the performance of our model by comparing it with various multi-modal foundational models as summarized in Table 1.

CLIP [15]. We fine-tune for our case using contrastive learning between image-text pairs. However, the CLIP model operates on small-sized images of 224×224 pixels. To address this limitation, we employ our proposed method of training of using ROIs to obtain visual embeddings from mammograms using CLIP vision encoder (V). The resulting image embedding, along with the textual embedding, undergoes contrastive loss learning. During testing, the model receives a mammogram and two prompts for binary classification.

$$\text{Prompts: } < \text{Cancer:\{Yes/No\}; Indication:\{Clinical History\}} > \tag{4}$$

We experiment with various CLIP configurations by switching vision encoders, and weight initializations as shown in Table 1.

LLava-1.5 [12]. We also fine-tune state-of-the-art foundational model LLaVa-1.5 using the LORA [8] method. Both the training and testing sets encompass complete mammogram images along with a prompt. The prompt comprises of patient's clinical history and a query regarding whether the patient has cancer or not. The prompts are formulated in a manner that elicits a binary output of either "Yes" or "No" from the model. The model undergoes fine-tuning specifically for a binary output of "Yes" or "No" as the expected output which is then used for binary classification. It is to be noted that during testing, the model does not output confidence/logit value for its prediction and hence we could not calculate the AUC metric for it.

4.2 Ablation Study

Table 2 illustrates the significance of each embedding within our network architecture. Notably, the inclusion of patient clinical history (text embeddings) significantly enhances the performance of our proposed module. Additionally,

[4] https://mammo-iitd-aiims.github.io/MMBCD

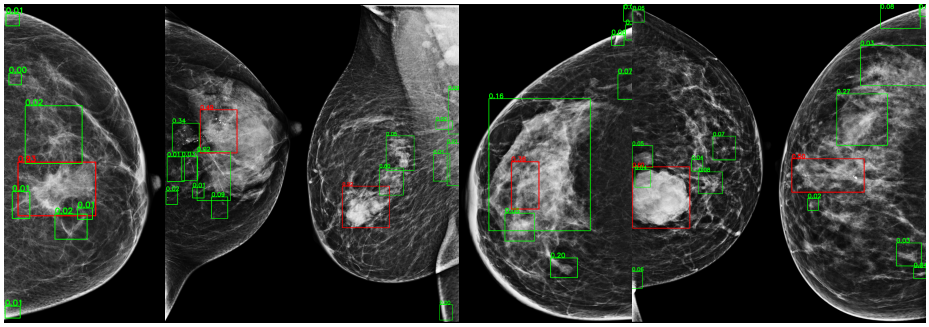

Fig. 3. The figure shows the ROI bounding boxes and the respective attention scores obtained from our proposed cross-attention layer for different samples. The ROI with the highest attention is drawn in red, and remaining in green.

Table 2. Performance impact of various components and embeddings in our model. "Attn" denotes attention module, and "mask" denotes use of history word masking.

				AIIMS 1			AIIMS 2		
Vision	**Text**	**Attn**	**Mask**	**Accuracy**	**F1-score**	**AUC**	**Accuracy**	**F1-score**	**AUC**
✓	-	-	-	0.94	0.69	0.928	**0.96**	0.66	0.900
✓	✓	-	-	**0.96**	0.79	0.968	0.94	0.60	0.943
✓	✓	✓	-	**0.96**	**0.82**	**0.973**	0.95	0.68	**0.950**
✓	✓	✓	✓	**0.96**	0.77	0.970	**0.96**	**0.71**	0.946

the effectiveness of our proposed cross-attention layer between text and visual embeddings is underscored by the attention visualization in Fig. 3, showcasing salient regions of mammograms receiving significant attention by the text embeddings. Note that for experiments without proposed attention module, we simply concatenate visual and text embeddings and pass it through an MLP for prediction. Furthermore, our approach prevents overfitting during text module training through proposed masking of specific words in the clinical history, as shown in results corresponding to "mask" in Table 2. In Table 3 we show another ablation study using different vision encoders in our architecture. Results of more ablation studies are in the supplementary.

5 Conclusion

Capturing clinical history of a patient straightforward, cost-effective, and holds immense value for radiologists, as it provides crucial context for interpreting imaging studies. Without pertinent clinical information, the significance of an imaging study may be diminished. Yet, most automated techniques for breast cancer detection do not include clinical history in the inference process. Our proposed approach addresses this key deficiency. In the process, we develop a novel architecture which leverages recent advances in multi-modal foundational models to give an highly accurate breast cancer detection.

Table 3. Ablation study by using various vision encoders in proposed model. As observed, `ViT-DINO` works best in our settings, and has been used in all experiments.

Model	Weights	AIIMS 1			AIIMS 2		
		Acc.	F1	AUC	Acc.	F1	AUC
ResNet50	ImageNet	0.94	0.63	0.914	0.96	0.62	0.882
ResNet50	CLIP	0.94	0.64	0.853	0.96	0.65	0.817
ViT	ImageNet	0.94	0.61	0.878	0.95	0.57	0.855
ViT	CLIP	0.94	0.66	0.908	0.96	0.65	0.884
ViT	DINO	0.94	**0.69**	**0.928**	0.96	**0.66**	**0.900**

Acknowledgments. We thank Mr. Rohan Raju Dhanakshirur for his assistance. We acknowledge and thank the funding support from Department of Biotechnology, India vide grant number BT/PR33193/AI/133/5/2019, and AIIMS Delhi-IIT Delhi Center of Excellence in AI funded by Ministry of education, government of India, Central Project Management Unit, IIT Jammu with sanction number IITJMU/CPMU-AI/2024/0002. Kshitiz Jain is supported by Yardi School of Artificial Intelligence, IIT Delhi via its Publication Grant for Students.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Burnside, E.S., Sickles, E.A., Bassett, L.W., Rubin, D.L., Lee, C.H., Ikeda, D.M., Mendelson, E.B., Wilcox, P.A., Butler, P.F., D'Orsi, C.J.: The acr bi-rads® experience: learning from history. Journal of the American College of Radiology **6**(12), 851–860 (2009) 3
2. Carney, P.A., Cook, A.J., Miglioretti, D.L., Feig, S.A., Bowles, E.A., Geller, B.M., Kerlikowske, K., Kettler, M., Onega, T., Elmore, J.G.: Use of clinical history affects accuracy of interpretive performance of screening mammography. Journal of clinical epidemiology (2012) 2
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 5
4. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence (1997) 2
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020) 2
6. Hager, P., Menten, M.J., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23924–23935 (2023) 2
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CVPR (2016) 2
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 7

9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. PMLR 2
10. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (Jan 2023), https://github.com/ultralytics/ultralytics 4
11. Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N.: Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis (2017) 2
12. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 7
13. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023) 2
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692 (2019) 4, 6
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 7
16. Rangarajan, K., Gupta, A., Dasgupta, S., Marri, U., Gupta, A.K., Hari, S., Banerjee, S., Arora, C.: Ultra-high resolution, multi-scale, context-aware approach for detection of small cancers on mammography. Nature Scientific Reports (2022) 2
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE TPAMI (2017) 4
18. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation **28**(1), 11–21 (1972) 6
19. Tang, Y., Cao, Z., Zhang, Y., Yang, Z., Ji, Z., Wang, Y., Han, M., Ma, J., Xiao, J., Chang, P.: Leveraging large-scale weakly labeled data for semi-supervised mass detection in mammograms. CVPR (2021) 2
20. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022) 2
21. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks. NeurIPS (2022) 4, 5
22. Zheng, H., Lin, Z., Zhou, Q., Peng, X., Xiao, J., Zu, C., Jiao, Z., Wang, Y.: Multitranssp: Multimodal transformer for survival prediction of nasopharyngeal carcinoma patients. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–243. Springer (2022) 2