Integrating Image Interpretation and Textual Context for Improved Breast Imaging Classification

Halil Ibrahim Gulluk
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
gulluk@stanford.edu

Olivier Gevaert
Biomedical Informatics Research (BMIR)
Stanford University
Stanford, CA 94304
ogevaert@stanford.edu

Abstract

Deep learning methods have demonstrated promising outcomes in predicting BI-RADS scores from mammography images. However, the interpretability of these images can vary, leading to discrepancies even among radiologists. Given the inherent complexity of mammography images, training classification models solely based on image labels often yields subpar performance. To overcome this challenge, we curated 2313 mammogram images and their corresponding captions from two mammography atlases. Our proposed approach employs a multi-modal model that leverages a pretrained PubMedBERT for the language component. By training this model on image-text pairs using contrastive learning, we empower our vision encoder to assimilate the rich information embedded within the captions, thereby enhancing its comprehension of mammography findings. Subsequently, we fine-tune the vision encoder using two datasets for BI-RADS prediction, achieving superior performance compared to models trained without pretraining, particularly when labeled samples are scarce. The enhancement in the 3-class average F1 score varies, ranging from +1% to +14%, depending on the number of training samples. Specifically, a +1% increase was noted when utilizing 40K training samples, while a +14% increase was observed with 1K samples. Furthermore, our experimental findings reveal that 2K image-text pairs from mammography atlases can be more informative than 2K labeled samples even for the label prediction, where the average margin is +1.1% when more than 10K training samples are present, which underscores the significance of incorporating textual information for modeling medical image data. As a result, our work provides a vision-language model for mammography and highlights the textual information from mammography atlases. The training code, pre-trained model weights, and data extraction scripts are publicly available at: https://github.com/igulluk/MAM-CLIP

1 Introduction

Breast cancer stands as a prominent cause of mortality across various cancer types. In 2020, the World Health Organization reported that 2.26 million individuals received a diagnosis of breast cancer [1]. Mammography, MRI, and ultrasound imaging techniques are commonly employed in both diagnosis and screening practices for breast cancer. Mammography, in particular, is widely utilized due to its rapid image capture capability, making it especially suitable for routine clinical evaluations. This study focuses on the development of models based on mammography images.

Mammography imaging involves capturing two primary views: the cranio-caudal (CC) and the mediolateral-oblique (MLO) views. Each patient undergoes four image captures in practice: LCC and LMLO for the left breast, and RCC and RMLO for the right breast. The findings in mammography

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: BI-RADS Scores and Corresponding Descriptions

BI-RADS Score	Diagnosis	Description
BI-RADS 1	Negative	No finding, normal
BI-RADS 2	Benign	Definite benign finding
BI-RADS 3	Probably Benign	Finding are benign with probability > %98
BI-RADS 0	Incomplete	Further information is needed for diagnosis
BI-RADS 4	Suspicious findings	Findings have possibility of malignancy (%3–%94)
BI-RADS 5	Highly suspicious of malignancy	Findings have possibility of malignancy (>%95)
BI-RADS 6	Positive	Biopsy proven malignancy

are typically reported using the Breast Imaging Reporting and Data System (BI-RADS) [2], which serves as the standard tool.

The BI-RADS system categorizes Mammography images into seven different classes to indicate the risk of breast cancer. The BI-RADS scores and their corresponding risks are detailed in Table 1. This study aims to predict BI-RADS scores from mammography images.

In recent years, researchers have focused on developing deep learning models for breast cancer detection and BI-RADS score prediction [3, 4, 5, 6]. In [4], authors utilized segmentation masks of breast lesions and selected small patches based on their intersection with the lesions. These smaller patches were then used to propose deep convolutional models for BI-RADS classification. Moreover, in [6], separate models were trained for BI-RADS and density classification for each view—LCC, LMLO, RCC, and RMLO. Subsequently, the LightGBM algorithm [7] was employed for final prediction.

In a separate study, authors proposed models for breast cancer detection by leveraging electronic health records in addition to mammogram images [8]. In [3], deep convolutional models were suggested for both small patch-based and whole-image classification of breast cancer using mammogram images. Furthermore, the combination of mammography images and clinical factors for estimating the malignancy of microcalcifications is proposed in [5].

Furthermore, researchers have been developing vision-language models in the medical domain at large [9, 10, 11, 12, 13]. These models have demonstrated utility either through fine-tuning for specific tasks or excelling at question answering using images. However, these works often utilize datasets with either very few or no mammography images. Moreover, the captions accompanying the images may lack detailed explanations of diseases, as they are not typically written for educational purposes. This underscores the value of our image-text dataset. In our study, we extract images and their captions from radiology atlases. These atlases are designed with the specific aim of educating radiologists, resulting in captions that are rich in information and meticulously explain details within the images to aid clinicians in accurate diagnosis.

In contrast to classical computer vision datasets like ImageNet [14], medical images often require more nuanced interpretation. This is particularly true for BI-RADS classification, where the interpretation by clinicians can be crucial. The presence of the BI-RADS 0 class, indicating a need for further information, adds to the complexity. Clinicians may assign a BI-RADS 0 label to images for various reasons, which could potentially have strong associations with other classes. To tackle this challenge, we developed a multi-modal model capable of accepting mammogram images along with corresponding captions from mammography atlases. This approach enables our models to capture complex information in breast images and potential reasons behind how and why images are labeled with specific BI-RADS scores. Our results demonstrate a significant improvement in BI-RADS classification through the integration of vision-language information.

2 Methodology

2.1 Vision-Language Model

Our objective is to train a vision-language model similar to the CLIP model [15], wherein images and their captions share similar representations, measured by cosine similarity. For the language component, we employ the PubMedBERT model [16] with pretrained parameters. For the visual

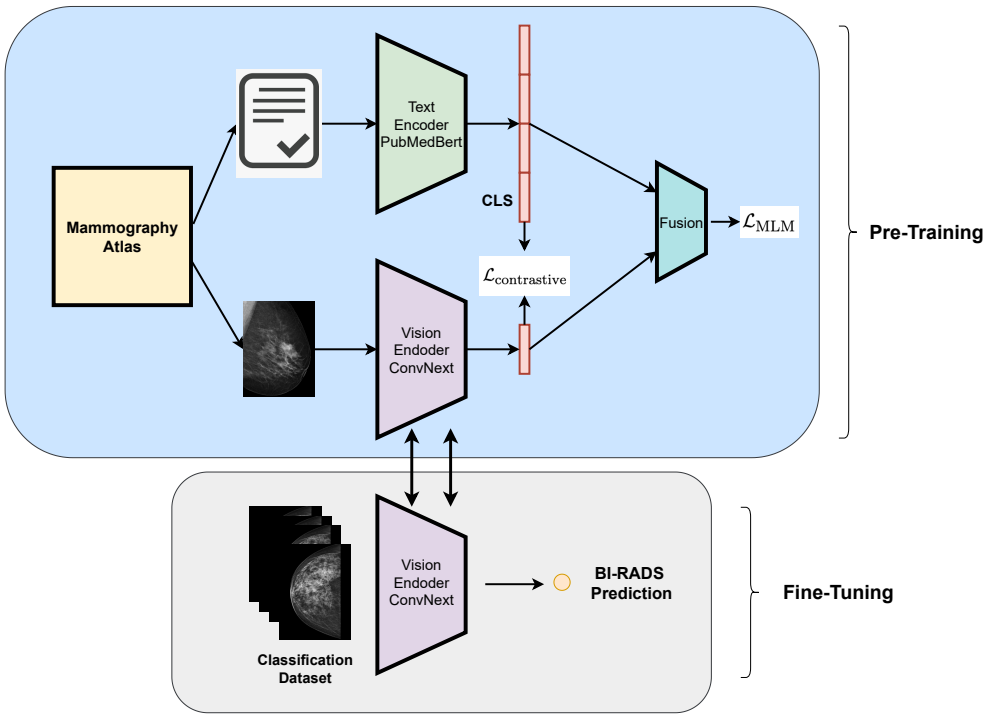


Figure 1: Model Overview. We first extract mammogram images and corresponding captions from Mammography Atlases and train a vision language model using contrastive and masked langauge modeling loss. Note that text encoder is pretrained PubMedBERT. Then we fine-tune vision encoder for our BI-RADS and density classification tasks on two datasets.

part, we utilize a ConvNext model [17] pretrained on ImageNet. Our observations indicate that using high-resolution mammography images is crucial for achieving better classification performance. Therefore, we opt not to use pretrained transformer-based vision encoders from Vision-Language Models such as Med-Flamingo [9] and BiomedCLIP [10], as they are trained on low-resolution images. Additionally, we find that ResNet-based models [18] exhibit relatively lower accuracy compared to ConvNext models. Consequently, we also refrain from using ResNet-based vision encoders from models like MedCLIP [11], PMC-CLIP [12], and PMC-VQA [13].

We adopt the training methodology outlined in PMC-CLIP [12], where the authors train the vision-language model using InfoNCE loss to enhance image-text pair similarity, along with masked language modeling (MLM) loss. The MLM loss helps in making the vision encoder as predictive as possible for masked language tokens.

To formalize, let's denote the visual encoder ConvNext model and PubMedBert model as Φ_{vis} and Φ_{text} respectively. Lets denote the current batch as follows:

$\mathcal{B} = \{(\mathcal{I}_1, \mathcal{T}_1), (\mathcal{I}_2, \mathcal{T}_2), ..., (\mathcal{I}_N, \mathcal{T}_N)\}$ where $(\mathcal{I}_i, \mathcal{T}_i)$ represents the ith image and text pair. Following the convention in [12] we denote the image representations by $\mathbf{v}_i = \Phi_{\text{vis}}(\mathcal{I}_i)$ and text representations by $\mathbf{T}_i = \Phi_{\text{text}}(\mathcal{T}_i)$ where $\mathbf{v}_i \in \mathbb{R}^d$, $\mathbf{T}_i \in \mathbb{R}^{l \times d}$. note that l stands for text length and d stands for the embedding dimension. Moreover, lets denote the [CLS] token representation by $\mathbf{t}_i \in \mathbb{R}^d$. We denote the cosine similarity between $\mathbf{v}_i$ and $\mathbf{t}_j$ by s_{ij} Then the contrastive learning loss becomes

$$\mathcal{L}_{\text{contrastive}} = \frac{-1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ij})} \right) - \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ji})} \right)$$

Moreover, in line with the masked language modeling approach described in [12], we randomly mask words with a specific probability and make predictions for those masked words using not only the text itself but also the vision embedding from the visual encoder. Specifically, we employ a transformer fusion model denoted as Φ_{fusion}, which takes the image embedding and masked text embedding to predict the ground truth text sequence. Let's denote the prediction of the fusion model for the masked

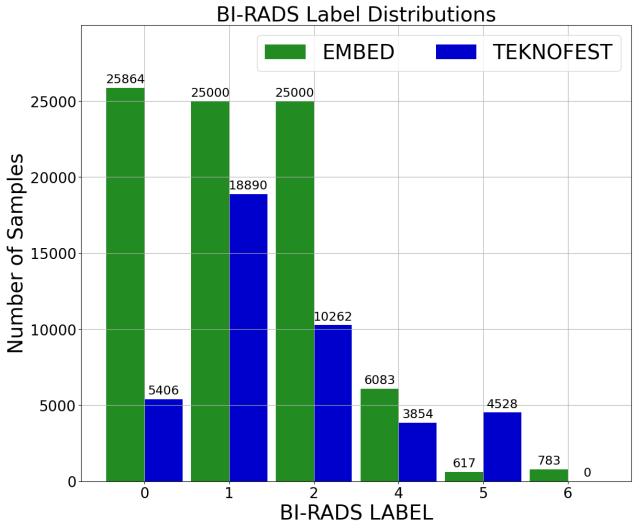


Figure 2: Number of patients for each classes in the classification datasets are provided. As it can be seen EMBED dataset is highly imbalanced in terms of BI-RADS labels. Also, there is no BI-RADS 6 patient in TEKNOFEST dataset.

token as $\mathbf{p}_i = \Phi(\mathbf{v}i, \mathbf{T}i^{\text{masked}})$. If the ground truth is $\mathbf{y}_i$, then the overall MLM loss can be formulated as follows:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{\mathcal{B}}[\mathcal{L}_{\text{CE}}(\mathbf{y}_i, \mathbf{p}_i))]$$

where $\mathcal{L}_{\text{CE}}$ is the cross entropy loss and the expectation is taken over the all masked tokens in the batch. Then with a specific choice of weight for the MLM loss λ, the overall loss for the pretraining can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{contrastive}} + \lambda \cdot \mathcal{L}_{\text{MLM}}$$

2.2 Pretraining Dataset Preparation

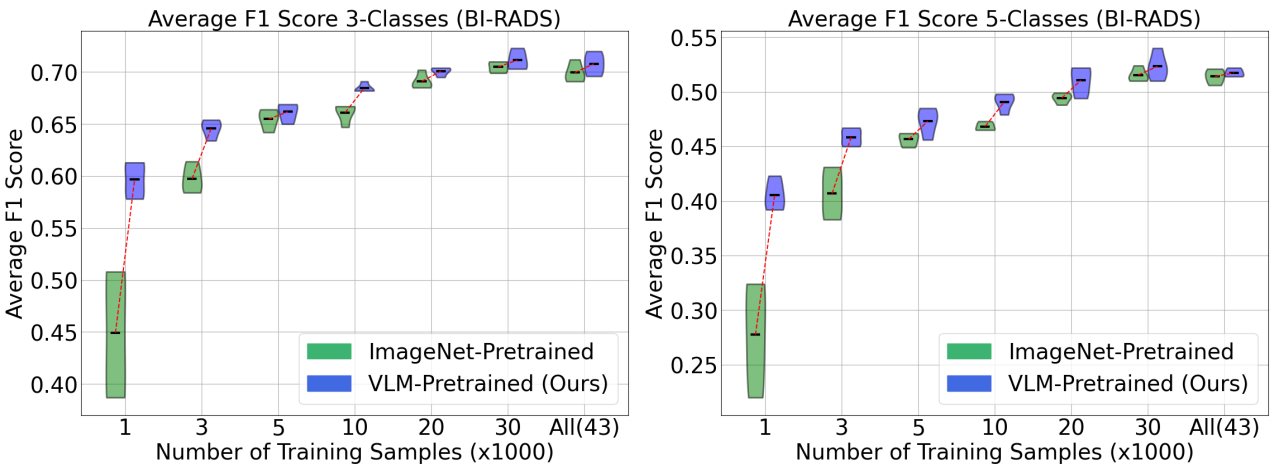
We extract mammography images and corresponding captions from two mammography atlases: Atlas of Mammography [19] and ACR BI-RADS ATLAS [20]. For Atlas of Mammography we used a python library named PyMuPDF [21] to extract the images and captions. On the other hand, we needed to use additional python library named pytesseract [22] to get the text from the images. If a caption describes more than one image, then we pair those images with the same caption as different image-caption pairs. Python scripts for extracting the image-text pairs are available in our code.

3 Classification Datasets

Our final goal is to predict BI-RADS class. To evaluate the performance of our models we utilize two different datasets: The Emory Breast Imaging Dataset (EMBED) [23] and TEKNOFEST, Artificial Intelligence in Healthcare Competition 2023 dataset [24].

EMBED: In the original dataset, there are different modalities than MLO and CC. However, we only utilize MLO and CC modalities. In addition, this dataset exhibits significant imbalance, as $\sim$%73 of the total images have BI-RADS 1 label. To address this imbalance, we limited the number of images with BI-RADS 1 or BI-RADS 2 label with 25,000 for each class. Furthermore, we exclude the BI-RADS 3 class from our experiments. The final label distributions are depicted in Fig. 2.

TEKNOFEST: TEKNOFEST dataset [25] was prepared for the Artificial Intelligence in Health Competition in 2023 [26] by the Republic of Turkey Ministry of Health. It comprises data from 10,735 patients, with four different images available for each patient: RMLO, RCC, LMLO, and LCC. Thus, a total of 42,940 images are included in the dataset, and the label distributions can be found in Fig. 2. The original DICOM files are scheduled to be made publicly available in 2024 by the Republic of Turkey Ministry of Health. Subsequently, preprocessed PNG files will be accessible upon request. Interested parties can obtain access to the data by contacting the corresponding author.



(a) (b)

Figure 3: TEKNOFEST dataset classification results are provided. For each number of training samples n (1,3,5,10,20,30,All(44) x1000) our vision-language pretraining outperforms ImageNet pretrained model and the difference between two models is very large for small number of training samples. Note that all experiments are done using 4-fold cross validation.

The preprocessing of DICOM images from both datasets mentioned above is performed using a YOLOX model [27]. This model is utilized to crop the breast from the background of the original DICOM images. Further details regarding this preprocessing step can be found in the code.

4 Experiments

We first train our multi-modal model using image-text pairs and select the best checkpoint in terms of validation loss. We then fine-tuned ImageNet pretrained models and the vision encoder of the Vision-Language Model (VLM) on two classification datasets. During the experiments, we merged images with a BI-RADS 6 label into the BI-RADS 5 class, treating them as if they had a BI-RADS 5 label. Additionally, we excluded the BI-RADS 3 class, which only exists in the EMBED dataset. Consequently, we worked with 5 different BI-RADS classes in the experiments. We conducted 4-fold cross-validation for all experiments.

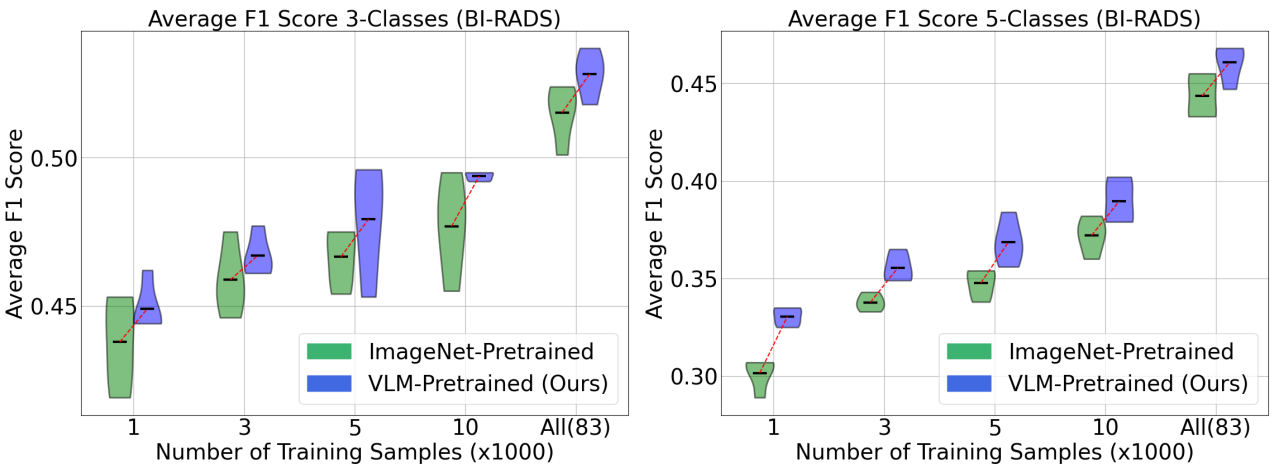
4.1 Implementation Details

For the visual and text encoders, we initialize an ImageNet pretraiend ConvNext [17] and pretrained PubmedBERT [16], respectively. Unlike many other multi-modal models, our image resolution is relatively high at 1024×768 pixels. We use a batch size of 64, AdamW optimizer [28] with a learning rate of 1e-4, and conduct training for 25 epochs. All experiments are performed using a single NVIDIA A100 GPU.

4.2 Results

For BI-RADS classification, we conducted a 5-class classification task. Additionally, we examined performance on a simplified set of 3 generic classes, where we grouped BI-RADS 1 and BI-RADS 2 together, as well as BI-RADS 4 and BI-RADS 5. Therefore, the final 3 classes are: BI-RADS 1,2 | BI-RADS 0 | BI-RADS 4,5. This formulation can be useful for clinical applications, as it allows us to interpret these classes as benign, needing further information, and malignant, respectively.

We computed class-based F1 scores and illustrated the average F1 scores for both 3 classes and 5 classes. The results for the TEKNOFEST and EMBED datasets are displayed in Fig. 3 and Fig. 4, respectively. Across both datasets, pretraining on mammogram images and texts led to an improvement in overall performance, even with varying numbers of training samples.



(a) (b)

Figure 4: EMBED dataset classification results. For each number of training samples, our vision-language pretraining outperforms ImageNet pretrained model. Note that all experiments are done using 4-fold cross validation.

Table 2: 3-Class Average F1 Performance on TEKNOFEST dataset, ImageNet-Pretrained models have 2000 more training samples in each case. For sample sizes larger than 10K, we observe that integrating textual knowledge with 2,313 samples provides more valuable insights than adding 2K labeled samples.

Model and # of Samples	Average F1	Average F1	Model and # of Samples
ImageNet-Pretrained 3K	**0.597±0.012**	0.597±0.015	VLM-Pretrained(Ours) 1K
ImageNet-Pretrained 5K	**0.654± 0.008**	0.646± 0.007	VLM-Pretrained(Ours) 3K
ImageNet-Pretrained 12K	0.668±0.008	**0.685±0.004**	VLM-Pretrained(Ours) 10K
ImageNet-Pretrained 22K	0.692±0.007	**0.701±0.003**	VLM-Pretrained(Ours) 20K
ImageNet-Pretrained 32K	0.697±0.006	**0.711±0.008**	VLM-Pretrained(Ours) 30K
ImageNet-Pretrained All(43K)	0.700±0.008	**0.706±0.002**	VLM-Pretrained(Ours) 41K

4.3 Assessing Textual Information: Captions vs. Labels

From the previous results, we observe an improvement in performance when using image-text pair pretraining. Furthermore, we experimentally demonstrate that captions accompanying images in the pretraining dataset can provide more informative cues than the labels of images in the classification dataset, despite the task being label prediction. In our pretraining dataset, comprising 2313 image-text pairs, we conducted experiments where the ImageNet-pretrained model had 2000 more data samples than our VLM-pretrained model for the classification task. As shown in Table 2, for cases where $n > 10000$, adding an additional 2000 samples did not yield as much improvement in results as pretraining with the original 2313 samples, even though the pretraining images come from a completely different source and distribution. This suggests that captions in mammography atlases can provide more guidance to deep models than labels of the images, owing to the complexities of mammography images.

5 Conclusion

In conclusion, we have demonstrated that explanations accompanying mammogram images, as found in mammography atlases, can play a crucial role in image classification tasks. Unlike other pretrained large models, leveraging only around 2300 image-text pairs from mammography atlases can lead to significant performance improvements, highlighting the informative nature of the captions in these atlases. Furthermore, our experiments indicate that these captions can often be more informative than the labels themselves. We have presented results for two new mammography datasets, underscoring the effectiveness of incorporating explanatory information in image classification tasks.

References

[1] World Health Organization. Cancer. `https://www.who.int/news-room/fact-sheets/detail/cancer`, Year of Access. Accessed: March 3, 2024.

[2] Laura Liberman and Jennifer H Menell. Breast imaging reporting and data system (bi-rads). *Radiologic Clinics*, 40(3):409–430, 2002.

[3] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):12495, 2019.

[4] Kuen-Jang Tsai, Mei-Chun Chou, Hao-Ming Li, Shin-Tso Liu, Jung-Hsiu Hsu, Wei-Cheng Yeh, Chao-Ming Hung, Cheng-Yu Yeh, and Shaw-Hwa Hwang. A high-performance deep neural network model for bi-rads classification of screening mammography. *Sensors*, 22(3):1160, 2022.

[5] Huanhuan Liu, Yanhong Chen, Yuzhen Zhang, Lijun Wang, Ran Luo, Haoting Wu, Chenqing Wu, Huiling Zhang, Weixiong Tan, Hongkun Yin, et al. A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of bi-rads 4 microcalcifications in breast cancer screening. *European Radiology*, 31:5902–5912, 2021.

[6] Huyen TX Nguyen, Sam B Tran, Dung B Nguyen, Hieu H Pham, and Ha Q Nguyen. A novel multi-view deep learning approach for bi-rads and density assessment of mammograms. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2144–2148. IEEE, 2022.

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[8] Ayelet Akselrod-Ballin, Michal Chorev, Yoel Shoshan, Adam Spiro, Alon Hazan, Roie Melamed, Ella Barkan, Esma Herzel, Shaked Naor, Ehud Karavani, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, 292(2):331–342, 2019.

[9] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[10] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[11] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[12] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023.

[13] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[16] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Ellen Shaw De Paredes. *Atlas of mammography*. Lippincott Williams & Wilkins, 2007.

[20] Mendelson EB Morris EA et al. D'Orsi CJ, Sickles EA. Acr bi-rads® atlas, breast imaging reporting and data system. reston, va, american college of radiology; 2013, 2013. `https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads` [Accessed: 2024/02/20].

[21] Jorj X. McKie. Pymupdf, 2024. `https://pymupdf.readthedocs.io/en/latest/module.html` [Accessed: 2024/02/20].

[22] Matthias Lee Lars Kistner Ryan Mitchell Emilio Cecchini Samuel Hoffstaetter, Juarez Bochi. pytesseract, 2024. `https://github.com/h/pytesseract` [Accessed: 2024/02/20].

[23] Jiwoong J Jeong, Brianna L Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilies, Geoffrey Smith, et al. The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 5(1):e220047, 2023.

[24] Artificial intelligence in health competition, teknofest. `https://teknofest.org/en/competitions/artificial-intelligence-in-health-competition/`. Accessed: 2024-03-03.

[25] T.C. Sağlık Bakanlığı. Mamografi verisi. `https://acikveri.saglik.gov.tr/Home/DataSetDetail/3`, 2024. Accessed: 2024-07-30.

[26] Teknofest. Teknofest: Aerospace and Technology Festival. `https://teknofest.org/en/`, 2024. Accessed: 2024-07-30.

[27] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.