Longitudinal Mammogram Risk Prediction

Batuhan K. Karaman, MS[1,2*], Katerina Dodelzon, MD[2], Gozde B. Akar, PhD[3], and Mert R. Sabuncu, PhD[1,2]

[1] Cornell University and Cornell Tech, New York, NY 10044, USA
[2] Weill Cornell Medicine, New York, NY 10021, USA
[3] Middle East Technical University, Ankara 06800, Turkey

Abstract. Breast cancer is one of the leading causes of mortality among women worldwide. Early detection and risk assessment play a crucial role in improving survival rates. Therefore, annual or biennial mammograms are often recommended for screening in high-risk groups. Mammograms are typically interpreted by expert radiologists based on the Breast Imaging Reporting and Data System (BI-RADS), which provides a uniform way to describe findings and categorizes them to indicate the level of concern for breast cancer. Recently, machine learning (ML) and computational approaches have been developed to automate and improve the interpretation of mammograms. However, both BI-RADS and the ML-based methods focus on the analysis of data from the present and sometimes the most recent prior visit. While it has been shown that temporal changes in image features of longitudinal scans are valuable for quantifying breast cancer risk, no prior work has systematically studied this. In this paper, we extend a state-of-the-art ML model [20] to ingest an arbitrary number of longitudinal mammograms and predict future breast cancer risk. On a large-scale dataset, we demonstrate that our model, LoMaR, achieves state-of-the-art performance when presented with only the present mammogram. Furthermore, we use LoMaR to characterize the predictive value of prior visits. Our results show that longer histories (e.g., up to four prior annual mammograms) can significantly boost the accuracy of predicting future breast cancer risk, particularly beyond the short-term. Our code and model weights are available at https://github.com/batuhankmkaraman/LoMaR.

Keywords: Breast Cancer Risk Prediction · Longitudinal Data · Transformer Neural Networks

1 Introduction

Breast cancer is one of the most diagnoses cancers worldwide [17], making population-level screening essential for early detection and improved treatment outcomes. Mammography (low-dose X-Ray imaging of the breasts) stands as the principal method among various breast cancer screening techniques, including

[*] Corresponding author: Batuhan K. Karaman, email: kbk46@cornell.edu

MRI and ultrasound, due to its accessibility, effectiveness, and reliability. Accurate risk prediction at the time of mammography is essential for adjusting screening intervals to suit individual patients and for determining the need for supplementary modalities for comprehensive evaluation. Traditional models for breast cancer risk assessment, such as the Tyrer-Cuzick (TC) and the Breast Cancer Surveillance Consortium (BCSC), utilize specific risk factors, including mammogram-derived expert-defined features such as breast density information, to estimate a patient's risk [10]. Yet, the reliability and utility of these predictions can vary. In response, recent advances have seen the emergence of deep learning-based algorithms that utilize mammographic data for breast cancer risk prediction. These approaches have shown promising results, often surpassing conventional risk models in their ability to predict both immediate and long-term risk of breast cancer [3,11,1,13,2,18,12,6].

However, the full potential of longitudinal data is often not realized in contemporary approaches. Most studies, including those using advanced deep learning techniques, predominantly analyze only the present mammogram, possibly together with limited prior visit, failing to take advantage of the comprehensive longitudinal mammographic history available. Despite this limitation, these models still achieve state-of-the-art prediction accuracy [2],[12]. Recently, [13] posited that a more extensive analysis that includes longer periods of screening data could lead to a richer understanding of the patterns and progression of breast cancer. Our paper aims to provide a systematic analysis of the utility of a broader range of historical data in quantifying breast cancer risk.

In this work, we present a Longitudinal Mammogram Risk model (LoMaR), which builds on [20] and implements a transformer architecture [16,5], coupled with a convolutional feature extractor [7], to ingest 2-view mammograms collected from an arbitrary number of (present and past) visits and predict breast cancer risk. We use a comprehensive dataset to evaluate our model and achieve state-of-the-art prediction performance, even with just the latest mammogram. Additionally, we explore how LoMaR benefits from the information provided by previous visits. Our analysis highlights the critical importance of longitudinal data in clinical environments, illustrating its critical role in enhancing the early detection of breast cancer.

2 Materials and Methods

2.1 Dataset

All participants in this study are from the Karolinska case-control (CSAW-CC) dataset [9], derived from the Cohort of Screen-Aged Women [4] in order to perform AI research to improve screening, diagnostics and prognostics of breast cancer. Karolinska dataset consists of women aged 40 to 74 who underwent mammography screenings over multiple years between 2008 and 2016. Each mammogram contains four X-ray scans from the craniocaudal (CC) and mediolateral oblique (MLO) views for both the left and right breasts, with all images captured using Hologic machines. The dataset includes a total of 19,328 mammograms, with 1,413 resulting in a cancer diagnosis, from 7,353 patients. Distribution of cancer-related labels and our image preprocessing steps are included in Sections S.1 and S.2 of the supplementary material, respectively.

2.2 Model

We focus on predicting an individual's future breast cancer diagnosis status using mammograms obtained from present and prior visits. To effectively harness the longitudinal sequence of visits and make future predictions for consecutive follow-up years after the present mammogram, we propose a transformer-based neural network which we refer to as LoMaR, which stands for Longitudinal Mammogram Risk model. LoMaR extends the state-of-the-art Mirai model that only ingests the present mammogram [20]. The detailed schematic of LoMaR is depicted in Figure 1.

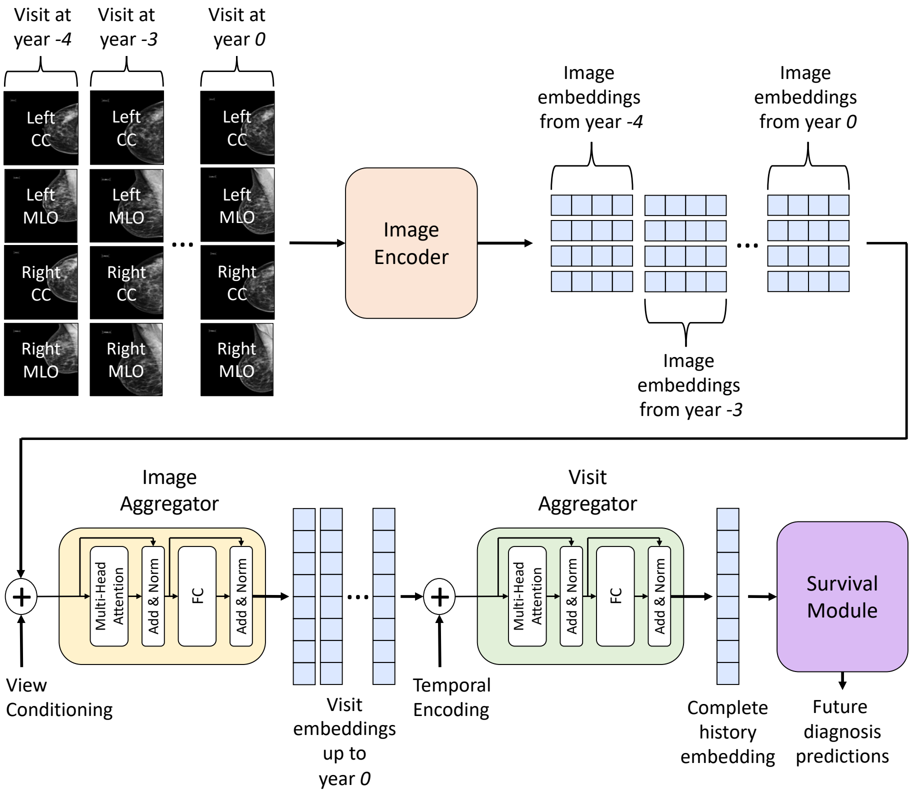

Fig. 1. Schematic representation of our Longitudinal Mammogram Risk model (LoMaR), where year 0 represents the current point in time.

LoMaR begins by processing each mammogram, comprising four 2D X-ray images of both left and right breasts in craniocaudal (CC) and mediolateral oblique (MLO) views, through a shared image encoder to produce distinct embeddings for each image. The image aggregator, a transformer encoder [5], then fuses the embeddings from the same visit, creating a visit embedding that integrates the information across the different views and both sides. These visit embeddings are then compiled by the visit aggregator, a subsequent transformer that processes each visit embedding as an individual token and captures the chronological progression of the visits. This results in a single, extensive embedding reflecting the entire longitudinal history. Finally, the survival module takes this embedding to predict the future cancer risk of the patient, employing a strategy that accounts for the progression of the disease over time.

In terms of network architecture, we utilize a ResNet-based convolutional neural network (CNN) [7] as the image encoder, followed by two transformer encoders that facilitate the aggregation of images and visits. The CNN and image aggregator employed in our model are identical to those described in [20]. More information about these two modules can be found in Sections S.3 of the supplementary material

The visit aggregator incorporates fixed sine and cosine positional encodings as temporal encodings at its input, as introduced in [16], which enables it to interpret the temporal ordering of visits. It processes the sequence through its self-attention block and employs average sequence pooling to distill the patient's entire mammographic history into a comprehensive embedding.

Our survival module is an additive hazard layer, similar to those found in [12] and [20]. It leverages the complete history embedding, denoted as m, for predicting future cancer risk. The module determines a baseline risk $B(m)$ and computes incremental risks for each subsequent year using linear layers. The incremental risk outputs are passed through ReLU activations. The cumulative k-year risk $P(t_{\text{cancer}} = k|m)$ is the sum of these risks, represented by the sigmoid of the sum of $B(m)$ and yearly hazards $H_i(m)$:

$$P(t_{\text{cancer}} = k|m) = \sigma(B(m) + \sum_{i=1}^{k} H_i(m)), \tag{1}$$

guaranteeing risk prediction growth over time.

2.3 Training

Despite their effectiveness, transformers are data-intensive and prone to overfitting due to their self-attention mechanisms; hence, we employed a training strategy to mitigate overfitting and improve model performance. We consider every visit within the training data as a present-time reference point, denoted as *present point*. For each visit identified as *present point*, we gather the patient's historical data, reaching back up to four years. This results in the most extensive training sequence comprising five visits: the visit four years before *present point*, three years before, two years before, one year before, and the visit coinciding with *present point*. We then track the subsequent diagnoses for the five following years after *present point*, creating a comprehensive 10-year window around each *present point* instance. By constructing these 10-year progression trajectories, we substantially expand the size of our training and validation datasets. To prevent information leakage, we ensure that all splits are at the individual subject level.

For the image encoder and image aggregator, we utilized the pretrained encoder and aggregator from [20], with both components frozen during training to retain their pre-learned features [19]. To address the imbalance in label distribution over the five-year follow-up horizon during training the visit aggregator and survival module, we use the reweighted cross-entropy loss function from [8] by adapting it for binary classification. The weights for each sample point are calculated based on the expanded datasets.

2.4 Evaluation

In evaluation, our main objective is to assess the influence of a patient's longitudinal mammographic history availability on the predictions made by our model. Notably, LoMaR is designed to handle incomplete histories, allowing for predictions with limited or no past data. We explore this by testing the model's performance across various history scenarios, simulating different durations and frequencies of patient history by selectively including or omitting visits. This methodology helps us understand the model's adaptability to real-world clinical settings where patient data may be incomplete or irregularly recorded.

In real-world longitudinal studies, biases in subject recruitment and follow-up are prevalent and can markedly affect the validity of the findings [21,15]. We employ an inference strategy designed to primarily mitigate the effect of temporal bias [22] in our results. We primarily use the area under the receiver operating characteristic (ROCAUC) metric, which accounts for the imbalance in the labels. To compute the ROCAUC score for a specific follow-up year (e.g, 2 years into the future) and longitudinal history scenario (e.g., with visits from present, and 1 year ago), we start by randomly selecting a single visit to represent *present point* for each test subject. This selection is made from all instances associated with a diagnosis in the designated follow-up year. The chosen instances of *present point* are then combined to form a 'random' pseudo test set. It is important to note that within this pseudo test set, each subject contributes a single follow-up diagnosis for prediction based on a specific historical instance, ensuring that all sample points in the pseudo test set are independent. After obtaining predictions using the visits that fit the specified longitudinal history scenario, we record the ROCAUC value. This process is repeated for multiple random pseudo test sets. Finally, we calculate the average of the ROCAUC values obtained from these pseudo test sets. This average ROCAUC score represents the comprehensive evaluation for the specific follow-up year and longitudinal history scenario pair. We note that the purpose of repeating the operation for multiple pseudo test sets is to broaden the range of disease progression trajectories we use in our evaluation and thus mitigate the potential effects of the aforementioned biases on our results.

3 Experiments

3.1 Implementation Details

We split the data into training and testing sets in an 80-20 ratio using a randomized, diagnosis-stratified approach, repeated 10 times. The results shown are the average of these iterations with 95% confidence intervals. Within each split, 25% of the training data was used for validation to determine early stopping based on validation loss. We optimized LoMaR's architecture using grid search across the splits, selecting the best architecture based on validation set performance. Fixed and tunable hyperparameters are detailed in Section S.4 of the supplementary material. ROCAUC calculations were performed using 100 random pseudo test sets.

3.2 Results

Table 1 presents C-index and ROCAUC scores for various models, including our LoMaR model tested across different longitudinal history durations and data collection frequencies, with Image-Only DL [18] and Mirai [20] as benchmarks. We calculate the C-index using the same method as we use for the ROCAUC. Image-Only DL and Mirai use the same image encoder and visit aggregator as LoMaR, while Mirai is considered as the state-of-the-art in the Karolinska dataset. Our proposed model, LoMaR, demonstrates superior performance over the Image-Only DL and Mirai models, even when no historical data are utilized, as shown by the ROCAUC scores across all follow-up years except follow-up year 2. For instance, with a ROCAUC score of 0.92 for the 1-year follow-up, our model significantly surpasses the Image-Only DL's score of 0.83 and Mirai's score of 0.90.

Impact of longitudinal mammographic history duration. Examining the ROCAUC scores for various longitudinal history durations in Table 1, we note that LoMaR shows a progressive improvement in long-term prediction performance as the history duration increases. For instance, when four years of past mammogram data are included, the 5-year ROCAUC score of LoMaR improves to 0.86, a substantial increase from the 0.81 achieved with no history ($\rho < 0.0001$). This improvement is consistent with the clinical understanding that a deeper historical context can provide more insight into the patient's breast health over time. It's important to also note that the earlier years' prediction performance remains stable regardless of the history duration included. This suggests that the initial changes are typically discernible in the mammograms at *present point*.

Table 1. C-index and ROCAUC scores of LoMaR and other existing models from the literature. LoMaR is evaluated by creating various longitudinal history durations (in years) and frequency scenarios in the test set. The notation * represents evaluations with annual data collection frequency, while † denotes evaluations with biennial data collection for the corresponding history duration.

	History		Follow-up year ROCAUC				
Model	duration	C-index	1-year	2-year	3-year	4-year	5-year
Image-Only DL [18]	0	0.75 (0.73–0.77)	0.83 (0.81–0.86)	0.79 (0.77–0.81)	0.75 (0.73–0.77)	0.73 (0.71–0.75)	0.71 (0.69–0.73)
Mirai [20]	0	0.81 (0.79–0.82)	0.90 (0.89–0.92)	0.86 (0.84–0.88)	0.82 (0.80–0.84)	0.80 (0.79–0.82)	0.78 (0.76–0.80)
LoMaR (Ours)	0	0.81 (0.78–0.83)	0.92 (0.90–0.94)	0.84 (0.83–0.86)	0.84 (0.83–0.85)	0.82 (0.81–0.83)	0.81 (0.80–0.82)
LoMaR (Ours)	$1^\star$	0.82 (0.78–0.83)	0.92 (0.90–0.94)	0.85 (0.83–0.86)	0.84 (0.83–0.85)	0.82 (0.81–0.83)	0.81 (0.80–0.82)
LoMaR (Ours)	$2^\star$	0.82 (0.78–0.83)	0.92 (0.90–0.94)	0.84 (0.83–0.86)	0.84 (0.83–0.85)	0.82 (0.81–0.83)	0.83 (0.82–0.84)
LoMaR (Ours)	$3^\star$	0.82 (0.78–0.83)	0.92 (0.90–0.94)	0.84 (0.83–0.86)	0.84 (0.82–0.85)	0.83 (0.82–0.84)	0.85 (0.84–0.86)
LoMaR (Ours)	$4^\star$	0.82 (0.78–0.83)	0.92 (0.90–0.93)	0.84 (0.83–0.86)	0.84 (0.83–0.86)	0.85 (0.84–0.86)	0.86 (0.85–0.87)
LoMaR (Ours)	$4^\dagger$	0.82 (0.78–0.83)	0.92 (0.90–0.93)	0.84 (0.83–0.86)	0.84 (0.83–0.86)	0.84 (0.84–0.86)	0.84 (0.85–0.87)

Impact of longitudinal data collection frequency. The comparison of longitudinal history durations $4^\star$ and $4^\dagger$ in Table 1 highlights the importance of more frequent screening. With an annual screening frequency ($4^\star$), the model achieves a 5-year ROCAUC score of 0.86, whereas with a biennial frequency ($4^\dagger$), the score is slightly lower at 0.84. This suggests that more frequent data collection can enhance the model's ability to predict long-term breast cancer risk, underscoring the value of annual screenings in improving risk assessment accuracy.

We observe a similar pattern of results when we exclude patients who have a cancer that was confirmed within 6 months of the *present point* (see Section S.5 of the supplementary material). In these results, notably, the performance boost of incorporating longitudinal history becomes even more pronounced.

Improved localization of cancer with past mammograms. In Fig. 2, we visualize Grad-CAM saliency maps [14] for the *present point* mammograms of a representative set of test subjects diagnosed with breast cancer. These subjects are examples of cases where historical mammograms enabled LoMaR to correct its initial prediction. The bottom row shows the expert-defined annotations of the cancer lesion. LoMaR, when given historical context, detects regions aligning with radiologist annotations for Subjects 0, 1, and 2, which were missed without history (second row). For Subjects 3 and 4, historical data narrows the model's focus and increases precision.

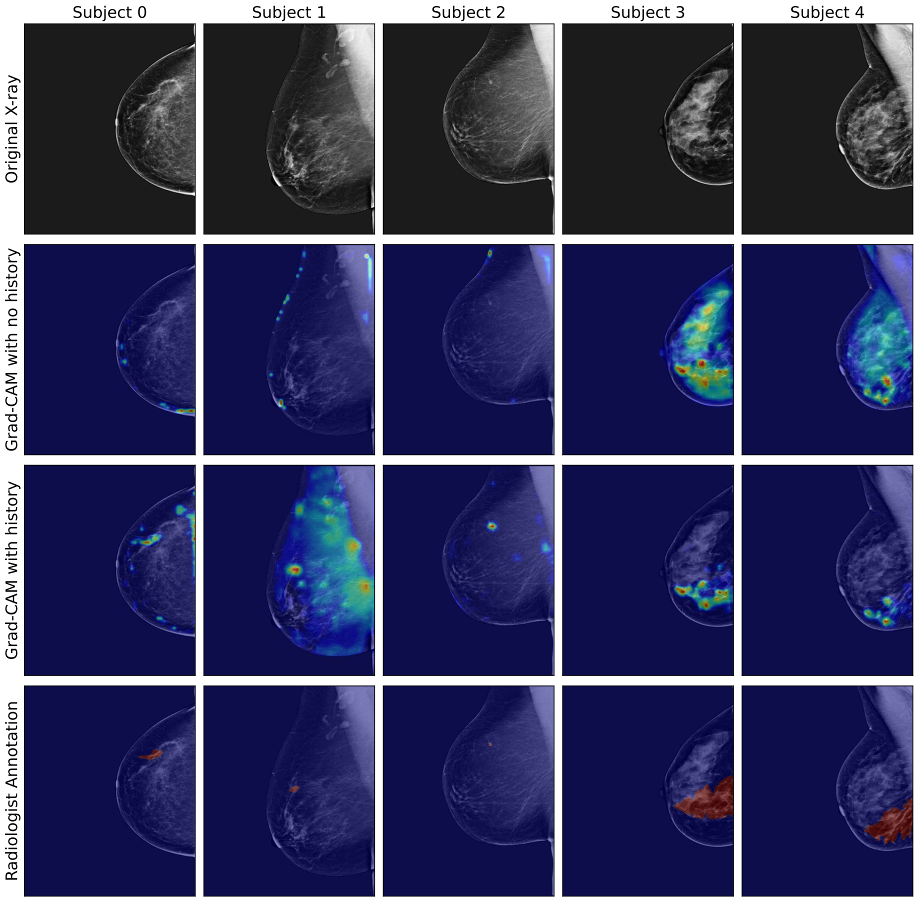

Fig. 2. Grad-CAM visualization of LoMaR with mammograms from five representative test subjects with (third row) and without (second row) prior mammograms.

4 Conclusion

In this paper, we extend a state-of-the-art ML model [20] to ingest an arbitrary number of longitudinal mammograms and predict future breast cancer risk. We use our model LoMaR to characterize the predictive value of prior mammograms. Our results show that longer histories (e.g., up to four prior annual mammograms) can significantly boost the accuracy of predicting future breast cancer risk, particularly beyond the short-term. This improved accuracy is due to the better localization of the suspicious lesion.

In the future, incorporating multimodal data can improve LoMaR's accuracy. Collaborations for clinical validation and robustness testing in real-world settings would be of significant interest. Additionally, exploring visualizations with methods alongside Grad-CAM can provide further insights.

Acknowledgement This research was in part funded by the National Institutes of Health (NIH) grant R01AG053949, and the National Science Foundation (NSF) CAREER 1748377 grant.

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

References

1. Arasu, V.A., Habel, L.A., Achacoso, N., Buist, D.S., Cord, J.B., Esserman, L.J., Hylton, N.M., Glymour, M.M., Kornak, J., Kushi, L.H., Lewis, D.A., Vincent, J.L., Lydon, C., Miglioretti, D.L., Navarro, D., Pu, A.X., Shen, L., Sieh, W., Yoon, H.C., Lee, C.: Comparison of mammography ai algorithms with a clinical risk model for 5-year breast cancer risk prediction: An observational study. Radiology **307** (06 2023). https://doi.org/10.1148/radiol.222733
2. Dadsetan, S., Arefan, D., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S.: Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. Pattern Recognition **132**, 108919–108919 (12 2022). https://doi.org/10.1016/j.patcog.2022.108919
3. Damiani, C., Kalliatakis, G., Sreenivas, M., Al-Attar, M., Rose, J., Pudney, C., Lane, E.F., Cuzick, J., Montana, G., Brentnall, A.R.: Evaluation of an ai model to assess future breast cancer risk. Radiology **307** (06 2023). https://doi.org/10.1148/radiol.222679
4. Dembrower, K., Lindholm, P., Strand, F.: A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks-the cohort of screen-aged women (csaw). Journal of Digital Imaging **33**, 408–413 (09 2019). https://doi.org/10.1007/s10278-019-00278-0
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 preprint arXiv:2010.11929 (2020)
6. Gastounioti, A., Desai, S., Ahluwalia, V.S., Conant, E.F., Kontos, D.: Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. Breast Cancer Research **24** (02 2022). https://doi.org/10.1186/s13058-022-01509-z
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (06 2016). https://doi.org/10.1109/cvpr.2016.90
8. Karaman, B.K., Mormino, E.C., Sabuncu, M.R.: Machine learning based multimodal prediction of future decline toward alzheimer's disease: An empirical study. PLOS ONE **17**, e0277322 (11 2022). https://doi.org/10.1371/journal.pone.0277322

9. Karolinska case-control dataset, https://data.europa.eu/data/datasets/https-doi-org-10-5878-45vm-t798?locale=en
10. Kim, G., Bahl, M.: Assessing risk of breast cancer: A review of risk prediction models. Journal of Breast Imaging **3**, 144–155 (02 2021). https://doi.org/10.1093/jbi/wbab001
11. Kim, H., Lim, J., Kim, H.G., Lim, Y., Seo, B.K., Bae, M.S.: Deep learning analysis of mammography for breast cancer risk prediction in asian women. Diagnostics **13**, 2247 (01 2023). https://doi.org/10.3390/diagnostics13132247, https://www.mdpi.com/2075-4418/13/13/2247
12. Lee, H., Kim, J., Park, E., Kim, M., Kim, T., Kooi, T.: Enhancing breast cancer risk prediction by incorporating prior images. Lecture Notes in Computer Science pp. 389–398 (01 2023). https://doi.org/10.1007/978-3-031-43904-9_38
13. Santeramo, R., Damiani, C., Wei, J., Montana, G., Brentnall, A.R.: Are better ai algorithms for breast cancer detection also better at predicting risk? a paired case–control study. Breast Cancer Research **26** (02 2024). https://doi.org/10.1186/s13058-024-01775-z
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
15. Tasci, E., Zhuge, Y., Camphausen, K., Krauze, A.V.: Bias and class imbalance in oncologic data—towards inclusive and transferrable ai in large scale oncology data sets. Cancers **14**, 2897 (06 2022). https://doi.org/10.3390/cancers14122897
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
17. Wilkinson, L., Gathani, T.: Understanding breast cancer as a global health concern. The British Journal of Radiology **95** (12 2021). https://doi.org/10.1259/bjr.20211033, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8822551/
18. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. Radiology **292**, 60–66 (07 2019). https://doi.org/10.1148/radiol.2019182716
19. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Satuluru, S., Kim, T., Banerjee, I., Gichoya, J., Trivedi, H., Lehman, C.D., Hughes, K., Sheedy, D.J., Matthis, L.M., Karunakaran, B., Hegarty, K.E., Sabino, S., Silva, T.B., Evangelista, M.C., Caron, R.F., Souza, B., Mauad, E.C., Patalon, T., Handelman-Gotlib, S., Guindy, M., Barzilay, R.: Multi-institutional validation of a mammography-based breast cancer risk model. Journal of Clinical Oncology (11 2022). https://doi.org/10.1200/jco.21.01337
20. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Smith, K., Wan, Y.L., Lamb, L., Hughes, K., Lehman, C., Barzilay, R.: Toward robust mammography-based models for breast cancer risk. Science Translational Medicine **13**, eaba4373 (01 2021). https://doi.org/10.1126/scitranslmed.aba4373
21. Yu, A.C., Eng, J.: One algorithm may not fit all: How selection bias affects machine learning performance. RadioGraphics p. 200040 (09 2020). https://doi.org/10.1148/rg.2020200040
22. Yuan, W., Beaulieu-Jones, B.K., Yu, K.H., Lipnick, S.L., Palmer, N., Loscalzo, J., Cai, T., Kohane, I.S.: Temporal bias in case-control design: preventing reliable predictions of the future. Nature Communications **12** (02 2021). https://doi.org/10.1038/s41467-021-21390-2