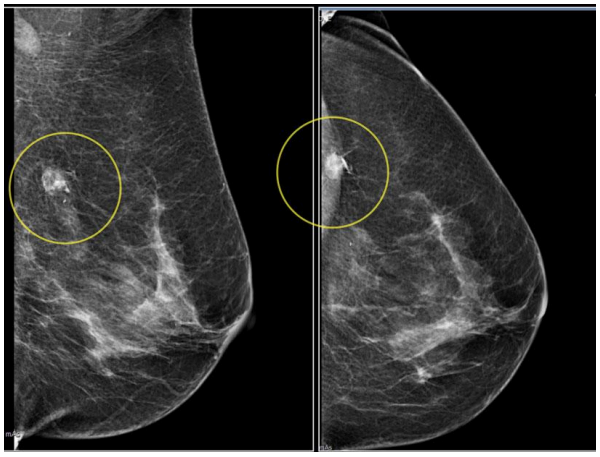
MA Thesis: Investigating Bias in AI Algorithms for Breast Cancer Detection from Mammography Imaging: A Focus on Generalization to Unseen Populations

Team:
Dilber Ozshahin[1], Charbel Mourad[2], Shadi Albarqouni[3,4]
[1] *College of Health Sciences, University of Sharjah, United Arab Emirates*
[2] *Lebanese Hospital Geitaoui HLG-CHU, Lebanese University, Lebanon*
[3] *Albarqouni Lab., Clinic for Interventional and Diagnostic Radiology University Hospital Bonn, University of Bonn, Germany*
[4] *Helmholtz AI, Munich, Germany*

Introduction. Breast density is a critical factor in breast cancer risk and detection, influencing the effectiveness of mammography. Higher breast density, characterized by a greater proportion of fibroglandular tissue relative to fatty tissue, is associated with a four- to sixfold increase in breast cancer risk. This risk is compounded by the fact that dense tissue can mask tumors on mammograms, reducing diagnostic sensitivity. Breast density typically decreases with age, particularly after menopause. However, this relationship is not uniform across all populations.



For instance, the study by Checka et al. (2012) [1] revealed that while 74% of women aged 40-49 had dense breasts, this percentage declined to 57% in women aged 50-59, and further to 36% in women aged 70-79. Notably, a significant portion of older women still possessed dense breasts, challenging the assumption that breast density diminishes uniformly with age. Similarly, Advani et al. (2021) [2] explored breast density among women aged 65 years and older, revealing that even in this older demographic, breast density remains a significant risk factor. The study found that 31.5% of women aged 65-74 had heterogeneously or extremely dense breasts, compared to 30.5% of those aged 75 and older. This persistence of high breast density in older women suggests that age alone is not a sufficient predictor of breast density.

Breast density and its impact on cancer detection vary significantly across different demographic groups. The studies highlight that certain racial and ethnic groups, such as Asian women, are more likely to have dense breasts, which may necessitate different screening protocols. Moreover, body mass index (BMI) also influences breast density, with lower BMI being associated with higher breast density, particularly in older women. For example, Advani et al. (2021) [2] found that 53.5% of women with heterogeneously or extremely dense breasts had a normal BMI, compared to 39.0% of women with scattered fibroglandular densities.

This project **aims** to investigate any potential biases in deep learning models trained predominantly on data from Western populations—primarily from Europe and the USA— and its generalization on data coming from different demographics. We **hypothesize** that these models may not adequately account for the variability in breast density across different demographic groups, leading to less accurate risk assessments and detection in non-Western populations.

The integration of deep learning (DL) models into mammography for breast cancer risk prediction has shown significant potential, yet there remains a notable **research gap** in understanding the implications of demographic variability on model performance. While existing models, such as the Tyrer-Cuzick model, have incorporated risk factors like breast density, these models have limitations, particularly when applied to diverse populations. Many DL models are trained on data from predominantly Western populations, which raises concerns about their generalizability and accuracy in non-Western or minority groups.

Moreover, the phenomenon of automation bias, as highlighted by Dratsch et al. (2023) [3], poses an additional challenge, particularly when AI systems are integrated into the radiology workflow. The risk of radiologists overly relying on AI suggestions without critical engagement is a significant concern, especially when the AI systems themselves may be biased due to the data they were trained on. This bias, coupled with the underrepresentation of diverse demographics in training datasets, underscores a critical gap in the current research landscape.

Given the complexities associated with breast density and its variation across demographics, it is crucial to develop and validate deep learning models that are sensitive to these differences. Such models should be trained on diverse datasets that reflect the global population, ensuring that they can accurately assess breast cancer risk and detect lesions in women from various demographic backgrounds. This approach will help mitigate the biases inherent in current models and improve the effectiveness of breast cancer screening worldwide.

Research Questions

Q1) How does the performance of deep learning models for breast cancer risk prediction vary across different demographic groups, particularly those underrepresented in the training data?
Q2) How do breast density and age affect the accuracy of DL-based breast cancer risk prediction models in non-Western populations?
Q3) How can AI-induced automation bias be mitigated in mammography to ensure equitable and accurate breast cancer screening across diverse populations?
Q4) What strategies can be employed to improve the generalizability of DL models in breast cancer risk prediction across various demographic groups?

Related works:
Recent studies have highlighted the growing role of deep learning in enhancing breast cancer risk prediction. Yala et al. (2019) [4] developed a DL model that outperformed traditional risk

models like Tyrer-Cuzick by leveraging full-field mammograms and traditional risk factors, achieving an AUC of 0.70 compared to 0.62 for the Tyrer-Cuzick model [4]. However, the model's performance varied across different demographic groups, particularly showing significant improvement in predicting risk for African American women compared to the Tyrer-Cuzick model. Dratsch et al. (2023) [3] focused on the impact of automation bias in AI-aided mammography. Their findings highlighted that radiologists, especially less experienced ones, might overly rely on AI suggestions, potentially leading to biased or inaccurate diagnoses. This underscores the need for careful integration of AI into clinical practice, ensuring that it complements rather than overrides human expertise. Lotter et al. (2021) [5] discuss the potential biases in deep learning models for breast cancer detection, particularly when trained on data that may not be fully representative of diverse populations. The authors emphasize the importance of evaluating models across different demographic groups to ensure robustness and fairness. They also address concerns about biases by validating their model on multiple datasets from different regions and institutions, demonstrating consistent performance across these diverse settings. Zufiria et al. (2022) [6] delve into the inherent biases present in mammography datasets that can impact the development and performance of deep learning models. The authors identify demographic biases, particularly related to age, race, and breast density, which can lead to uneven model performance across different population groups. They stress the importance of curating diverse and representative datasets to improve the fairness and generalizability of AI models in mammography. Readers are referred to Azam et al. (2018) [7] for more detailed papers on Breast Mammography.

Datasets. The landscape of publicly available mammography datasets is rich with resources critical for advancing breast cancer research, particularly in developing deep learning models. INbreast offers 410 images **Portugal** with comprehensive annotations, including lesion contours and breast density. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) refines DDSM data into 3,000 high-quality mammography images from the **USA**. The Mammographic Image Analysis Society (MIAS) Database contains 322 images from the **UK**, categorized by breast density and abnormality type. Lastly, the OPTIMAM Mammography Image Database (OMI-DB) is a vast repository with over 100,000 mammograms from the **UK**, rich in BI-RADS data and lesion annotations, serving as a vital resource for AI model training.

Besides, we have access to a couple of databases from different demographies, namely, VinDr-Mammo, from **Vietnam**, which consists of 5,000 four-view exams with breast-level assessment and finding annotations, and The Chinese Mammography Database (CMMD) which consists of 1,775 patients from **China** with benign or malignant breast disease, and KAUMD, from **Saudi Arabia,** which provides around 1500 cases with a total of 5600 mammogram images with BI-RADS scores. Besides, through our clinical partners, we have access to 1500 cases from **Lebanon** and around 100 cases from the **United Arab Emirates**. All cases have at least a Mammography scan for both left and right sides with multiple views.

The BI-RADS scores are reported and Malignant cases are biopsy proven. These datasets are indispensable for developing robust, generalizable AI models, though they vary in the level of demographic detail provided.

Roadmap (6 months):

- Familiarize yourself with the current literature [3-6]
- Build the baseline supervised models.
- Run the necessary comparisons.
- Run extensive experiments and analysis
- Write up your thesis

Requirements:

- Solid background in Machine/Deep Learning
- Familiar with deep learning models and SOTA architectures
- Sufficient knowledge of Python programming language and libraries (Scikit-learn)
- Experience with a mainstream deep learning framework such as PyTorch.
- Machine/Deep learning hands-on experience

References:

1. Checka, Cristina M., et al. "The relationship of mammographic density and age: implications for breast cancer screening." *American Journal of Roentgenology* 198.3 (2012): W292-W295.
2. Advani, Shailesh M., et al. "Association of breast density with breast cancer risk among women aged 65 years or older by age group and body mass index." *JAMA network open* 4.8 (2021): e2122810-e2122810.
3. Dratsch, Thomas, et al. "Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance." *Radiology* 307.4 (2023): e222176.
4. Yala, Adam, et al. "A deep learning mammography-based model for improved breast cancer risk prediction." *Radiology* 292.1 (2019): 60-66.
5. Lotter, William, et al. "Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach." *Nature medicine* 27.2 (2021): 244-249.
6. Zufiria, Blanca, et al. "Analysis of potential biases on mammography datasets for deep learning model development." *International Workshop on Applications of Medical AI*. Cham: Springer Nature Switzerland, 2022.
7. Hamidinekoo, Azam, et al. "Deep learning in mammography and breast histology, an overview and future trends." *Medical image analysis* 47 (2018): 45-67.
8. Muthukrishnan, Ramya, et al. "MammoDL: mammographic breast density estimation using federated learning." *arXiv preprint arXiv:2206.05575* (2022).
9. Schmidt, Kendall, et al. "Fair evaluation of federated learning algorithms for automated breast density classification: The results of the 2022 ACR-NCI-NVIDIA federated learning challenge." *Medical Image Analysis* 95 (2024): 103206.

If you are interested in this project, please write an email to shadi.albarqouni@ukbonn.de