



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/239119/>

Version: Accepted Version

Proceedings Paper:

APPIAH, KOFI ESSUMING (2026) Fairness in Breast Cancer Diagnosis with Deep Learning. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). , pp. 7350-7356.

<https://doi.org/10.1109/ICCVW69036.2025.00763>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Fairness in Breast Cancer Diagnosis with Deep Learning

Laura Mata Le Bot
Department of Computer Science,
University of York
York, YO10 5GE, UK
lmlb500@york.ac.uk

Sze Chai Kwok
Jiangsu Provincial University Key (Construction) Laboratory
For Smart Diagnosis and Treatment of Lung Cancer,
Duke Kunshan University, Kunshan, Jiangsu, China
szechai.kwok@duke.edu

Emmanuel Amankwaa-Frempong
Oncology Department,
Sweden-Ghana Medical Center
Accra, Ghana
emmanuel.frempong@sgmcltd.com

Kofi Appiah
Department of Computer Science,
University of York
York, YO10 5GE, UK
kofi.appiah@york.ac.uk

Abstract

This paper aims to explore the effectiveness of machine learning models for the classification of breast cancer images trained with data from a demographic region and tested or used in a different demographic region. Artificial intelligence has the potential to be integrated into the imaging process to reduce workload and broaden screening audiences. However, artificial intelligence has sometimes been shown to demonstrate bias in medical applications. Most available breast cancer images are collected in North American, European, or East Asian countries, and there is limited data available from other regions. Bias between demographics could lead to some groups being under-diagnosed, resulting in worsened prognoses. In this work, a high-performance breast cancer classification model with AUC of up to 0.7415 on ultrasound images and 0.8920 on mammogram images has been developed. The model is trained and tested on a variety of datasets, some specifically collected in Ghana to compare with publicly available datasets from the UK, Portugal, and Poland. Experiments were conducted to determine any bias between different demographic regions. A significant decrease in performance was found in five of the six experiments conducted.

1. Introduction

Breast cancer is the second most frequently diagnosed cancer world-wide, responsible for 11.6% of all cancers globally [3]. Breast cancer screening and diagnosis is reliant on medical imaging, most commonly mammography and

ultrasound. Mammography is the recommended form of imaging for screening. However, it is not suitable for higher density breasts, particularly in younger people or some ethnic groups. [18]. Ultrasound is typically used to diagnose patients with symptoms or signs of cancer in screening.

Breast cancer imaging requires access to equipment and highly-trained radiologists. Interpreting images requires a large amount of time and a high-level of human focus. Results are also subject to human errors which can delay diagnosis and treatment with serious consequences for the patient. Artificial intelligence (AI) has potential to be integrated into imaging processes and help to overcome these challenges. For example, machine learning (ML) can be used as an additional image reader, real-time assistant for radiologists, or even as an automated single reader for screening or tumour classification [4]. The use of AI in breast cancer imaging could reduce the workload of radiologists, reduce the risk of errors, and widen access to breast cancer screening.

There is evidence of bias in AI in many fields including the medical field. Algorithms can reflect the bias of their training data, which reflects human and societal biases. When AI is used for medical purposes, bias can have serious consequences. It has been shown that AI for dermatological diagnoses can under-diagnose certain ethnicities [1], and that medical forecasting models may perform better for patients with better access to healthcare [16]. Bias in AI may present differently across different areas in medicine. There is limited research into the causes and effects of bias in AI for breast cancer imaging. There is a potential for factors like ethnicity, genetics, healthcare access, or lifestyle factors to cause images to present differently and introduce bias into imaging AI.

Information about characteristics that could cause bias, like ethnicity, is not included in public imaging datasets. However, these characteristics do tend to differ across demographic regions. Studies demonstrating the successful use of AI for breast cancer imaging are typically trained with images from only North American, European, and East Asian datasets. There has been little investigation into how these models perform in different regions. Additionally, it is possible that there is further bias contained within datasets, for example, some ethnic groups may be underrepresented. The aim of this work is to investigate the presence of bias in a breast cancer classification model across demographic regions. Breast cancer images will be used from two different demographic regions: Sub-Saharan Africa (Ghana) and Europe (UK, Portugal, and Poland).

2. Related Work

Machine learning for breast cancer can be integrated into healthcare for practical use in a number of ways. Models may be deployed as a second or third reader, where both a radiologist and the model interpret the image and the results are compared [5]. Models may also be used as real-time assistants that flag certain images or areas within images as higher-risk for malignancy [6]. Finally, breast cancer classifiers may be deployed as single automated readers. This means that images are interpreted only by AI, with no radiologist input. This could potentially enlarge breast cancer screening audiences, however, the bioethical implications are not yet regulated [4].

One study [5] investigated whether AI cancer-detection software, INSIGHT MMG [13] could be used to triage mammograms into workstreams of no further radiologist assessment or enhanced assessment. The study [5] used up-sampling of healthy women to create a screening-like cohort and mitigate the effects of a cancer-enriched datasets. The triage was found to halve radiologist workload without missing cases.

This research was later expanded by assessing for non-inferiority of AI involvement in breast cancer screening [6]. The INSIGHT MMG [13] software was used again as a standalone classifier and an assistant for single and double radiologists' readings. This was compared to unassisted double radiologist readings with radiologists of varying skills. Overall, the study found that assisted single reading and the standalone classifier were non-inferior compared with unassisted double readings. [6]

Bias in AI may be present due to prevalence, presentation, or annotation disparities [12]. Prevalence disparities appear due to subgroups having different levels of prevalence in the data. Presentation disparities occur when different subgroups present with a disease differently. Annotation disparities are introduced by human individuals labelling data using different criteria. Some of these dispari-

ties could be derived from legitimate biological differences and should therefore be used by the model to inform predictions. In other cases, disparities may be caused by historical bias and must be mitigated by the model.

Machine learning models for healthcare applications can be limited by the generalisability of the models. Models tend to be trained with data collected from one healthcare centre, with the model later tested and deployed in other healthcare centres. However, several studies have found that there is a gap between locally reported model performance and cross-site generalizability [24]. This means that the models do not perform as well when tested with data from centres different to their training data.

Breast cancer classification models can exhibit bias caused by the prevalence of classes. Breast cancer datasets typically contain images from normal, benign, and malignant classes. Often, these datasets are cancer-enriched, which means that they contain a higher proportion of malignant samples than would be found in a general population. A breast cancer classification model may contain biases towards the most prevalent class in its training data [7]. This can cause a difference in performance when the model is tested with data with a different proportion of classes, for example, screening of the general population. Such biases can be mitigated with a variety of strategies, for example, by choosing to over-sample certain classes in a dataset [7].

The presence of bias in AI for dermatology is a widely discussed issue [1] as the cause of the bias and its consequences are clear. AI has the potential to be used for the diagnosis of skin conditions without the need for pathological tests [1]. The available datasets for the diagnosis of skin conditions contain a higher volume of images of White skin. However, skin conditions appears very different for different ethnicities. A machine learning model trained on the available datasets will learn the biases of the datasets, and be more likely to correctly diagnose skin cancer in White patients than patients of other ethnicities [1].

A scoping review of racial bias in clinical machine learning found racial bias present in eight out of ten studies [11]. One study analysed underdiagnosis of ethnic minorities in chest X-ray classification models [20]. Underdiagnosis is defined as falsely claiming that a patient is healthy. A model was trained on a complete dataset. The test data was divided into subgroups, where some subgroups were known to be underserved. The model used a DenseNet architecture pre-trained with ImageNet. The results showed that patients who were female, under 20 years old, Black or Hispanic, or of lower socio-economic status were at higher risk of underdiagnosis. Belonging to more than one underserved group also increased the risk.

Another study analysed bias in breast cancer survival models [16]. Several models were trained on specific ethnicity and compared to generic models. Four architectures,

Cox Proportional-Hazards (CoxPH), Gradient Boost Tree (GBT), Survival Tree (ST), Survival Support Vector Machine (SSVM) were used. The models were evaluated using a standard performance measure for survival models called C-index. The results showed that ethnicity-specific models consistently out-performed general models for the task of predicting survival.

This review has identified potential for bias between demographic regions in machine learning for breast cancer imaging. This bias could be caused by limited prevalence, where some demographic regions are underrepresented in the available training datasets. There may also be presentation disparities, where certain ethnicities present differently in breast cancer images. Finally, it is possible that annotation disparities are present, with images being labelled according to different criteria across demographic regions or healthcare centres. The study demonstrating the development of INSIGHT MMG [13] indicated that there may be a decrease in performance when a model is trained and tested with data from a different demographic region. Further research should be conducted into the presence of this bias, but may be limited due to models being insufficiently generalisable [24].

3. Our Approach

This section outlines the approach taken in this work. Suitable datasets are assessed and selected. The steps for data pre-processing are described. The model architecture is selected and the implementation and validation strategies are outlined. The bias analysis experiments are planned.

3.1. Dataset Selection

The aim of the work is to train a model using images from one demographic region then test the model with data from another region. The idea was to use ultrasound and mammography datasets collected specifically for this purpose from two different continents; Africa and Europe, with each set containing a single ethnicity. The dataset from Europe will be substituted with publicly available datasets collected in Europe. The criteria below was used to identify the most suitable public datasets:

1. **Relevance to goal:** The datasets must be suitable for investigating demographic bias. Each dataset must have been collected in a single demographic region.
2. **Size and quality:** The datasets must contain sufficient images to train and test the model. The datasets can be cancer-enriched and contain a high proportion of malignant samples.
3. **Labels:** The images must be labelled by expert radiologists. Each image should be classified as normal, benign, or malignant.
4. **Legal and ethical compliance:** The datasets must comply with legal and ethical guidelines and have full ethical

approval to train machine learning models.

5. **Comparability of models:** The datasets should be of similar size and have similar class prevalence to allow models to be compared.

The Sweden Ghana Medical Centre has provided ultrasound and mammogram images of exclusively Ghanaian women (Ghana). Mammography datasets from regions of predominantly White ethnicity used are listed in Table 1. The only publicly available dataset collected in the UK is MIAS [21], which is comprised of digitised film images. The INBreast dataset [15] is also publicly available and contains digital images, however, it was collected in Portugal, not the UK. Figure 1 shows a sample image from each of these datasets. The MIAS image is clearly of lower quality than the Ghana and INBreast images. It was decided to conduct mammogram experiments using both datasets. Although, there is no guarantee that the publicly available datasets (MIAS and INBreast) contain a single ethnicity, but they were both collected in European countries of predominantly White ethnicity.

Dataset	Region	Availability	Description
INBreast	Portugal	Public	410 digital images
MIAS	UK	Public	322 film images

Table 1. list of EU mammography datasets.

There is limited availability of public ultrasound datasets. The literature reviewed identified two possible datasets: BrEaST [17], collected in Poland, or BUS-BRA [9], collected in Brazil. The BrEaST dataset was selected as it was collected in a country of predominantly White ethnicity.

3.2. Dataset Descriptions

The selected datasets contain labels for normal, benign, and malignant cases. The region, image type, and number of samples present for each class is outlined in Table 2. The BrEaST dataset uses BIRADS scores [22] rather than normal, benign, and malignant classes. These have been mapped to the most likely corresponding class. BIRADS score 1 is mapped to normal, scores 2 and 3 are mapped to benign, and scores 4 and 5 are mapped to malignant [22].

Name	Type	Normal	Benign	Malignant
Ghana	Ultrasound	192	178	167
BrEaST	Ultrasound	4	154	98
Ghana	Mammogram	255	360	299
MIAS	Mammogram	207	64	51
INBreast	Mammogram	67	243	100

Table 2. Description of samples in the selected datasets.

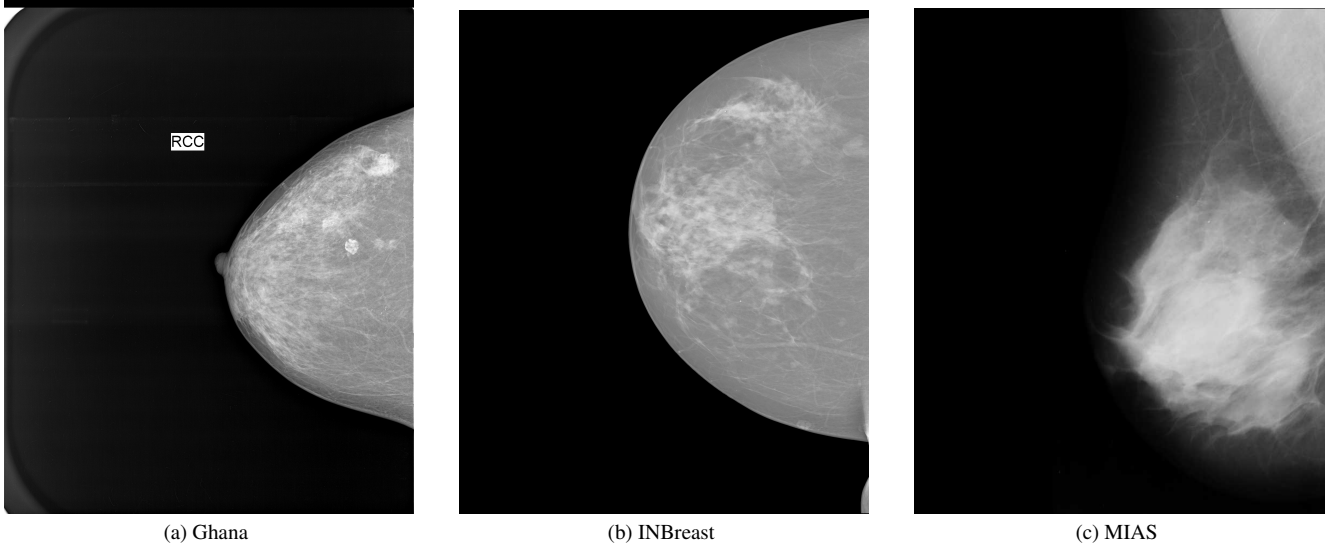


Figure 1. Sample images from the selected mammogram datasets.

3.3. Dataset Pre-Processing

In the pre-processing stage, the datasets will be divided into training, validation, and test sets. The model will be tested on two different test sets. The test sets initially contain a different number of samples from each class. This could cause a change in performance if the model is better at predicting certain classes. The experiments look for a change in performance caused by demographic regions and causes like class imbalance should be mitigated. The test sets will be balanced to contain exactly the same number of samples from each class. This is achieved by taking the minimum number of samples available per class in either set. For example, if one test set contains 50 malignant samples and the other contains 40, the balanced test sets will contain 40 malignant samples each.

3.4. Model Selection

The experiments will each be conducted using the same architecture. The model should perform highly at the breast cancer classification task and train relatively quickly, as the experiments will require up to 50 iterations. The literature review identified two possible architectures for the classification model. The INSIGHTMMG software, used in several studies, makes use of the ResNet architecture and has been shown to be non-inferior to the work of trained radiologists [13]. The VGG architecture has been shown to perform very highly for both mammograms and ultrasounds [19][10]. Both of the identified architectures are available in Python and Tensorflow with capacity for transfer learning. From the Literature Review, it is expected that VGG could perform highly and consistently for both ultrasound and mammograms. Both architectures have been tested dur-

ing the initial development phase before selecting the final architecture.

3.5. Model Validation

Performance metrics will be used to evaluate the effectiveness of the model. These performance metrics expect that the prediction made by a model is a true positive, false positive, true negative, or false negative. In this case, the model will perform multiclass classification and predict normal, benign, or malignant labels, which are not divided into positive and negative. To compute performance metrics over the three classes, a One-vs-Rest (OvR) strategy will be applied [8]. Each label will be treated as the positive class in turn and the metric will be computed for that label. The macro-average of the three metrics will be taken and used as the final metric. The goal of the model is to correctly classify breast cancer images as malignant, benign, and normal. A false negative that incorrectly classifies a malignant image as normal or benign has serious consequences. The AUC, computed with the OvR strategy, is the primary metric identified in the Literature Review to assess performance of cancer classification models, and will be used to compare the performance of the models.

3.6. Bias Analysis

In this paper, the term bias refers to the performance of the classification model changing between different datasets. The goal is to detect if bias is present between demographic regions and to determine the cause of this bias. The experiments will look for a change in performance when the model is tested with data from a different demographic region to its training data. For these experiments, the Ghana refers to the dataset collected in Ghana.

All of the experiments will follow the same structure. One dataset (the primary dataset) will be split into training, validation, and test sets. A second test set will be obtained from a different dataset mainly from another demographic region and referred to as the secondary dataset. The model will first be trained with the primary training set. The model will then be tested with the primary test set and then the secondary test set separately. The datasets that will be used for each experiment are outlined in Table 3. The performance of the model on the primary and secondary test set of each experiment will be compared.

Experiment ID	Image type	Primary dataset	Secondary dataset
DM-1A	Mammogram	Ghana	MIAS
DM-1B	Mammogram	MIAS	Ghana
DM-2A	Mammogram	Ghana	INBreast
DM-2B	Mammogram	INBreast	Ghana
DU-1A	Ultrasound	Ghana	BrEaST
DU-1A	Ultrasound	BrEaST	Ghana

Table 3. List of bias experiments and datasets.

The results of the experiments will be analysed to determine whether there is a change in performance across the two test sets. The following null hypothesis will be tested:

H_0 : There is no performance change between the classification model tested with data from the same demographic region or a different demographic region to its training data. The experiment will run for 50 iterations to ensure sufficient samples are generated for statistical analysis. AUC will be used to perform statistical test as it is the main metric used for comparing cancer classification model performances in the Literature Review. The two test sets are derived from independent datasets, therefore the AUCs for each iteration are also independent. The distribution of the AUCs for each dataset over the 50 iterations can be compared with a Student's t-test or a Mann-Whitney test [14]. The independent Student's t-test assumes that data is normally distributed, so the distributions will first be tested for normality using the Shapiro-Wilk test [23]. For Shapiro-Wilk, the null hypothesis H_0 will be that the AUCs follow a normal distribution. If H_0 for Shapiro-Wilk is accepted for both test sets, the Student's t-test will be used to test for a change in performance. If H_0 for Shapiro-Wilk is rejected for either test set, the Mann-Whitney test should be used instead.

4. Results

The datasets used for each experiment is shown in Table 3. The experiments tested for a change in performance between datasets from different demographic regions. For each experiment, the model was trained and tested 50 times. The accuracy and AUC were recorded for the primary and

secondary test sets. The mean value of each metric across the 50 iterations is recorded in Table 4 for the primary and secondary test sets in each experiment.

ID	Image type	Prim. Test Set	Sec. Test Set	Mean ACC	Mean AUC
DM-1A	Mamm	Ghana	Ghana	0.66	0.89
		Ghana	MIAS	0.41	0.53
DM-1B	Mamm	MIAS	MIAS	0.51	0.52
		MIAS	Ghana	0.56	0.51
DM-2A	Mamm	Ghana	Ghana	0.64	0.83
		Ghana	INBreast	0.38	0.53
DM-2B	Mamm	INBreast	INBreast	0.67	0.72
		INBreast	Ghana	0.32	0.54
DU-1A	USG	Ghana	Ghana	0.86	0.74
		Ghana	BrEaST	0.63	0.53
DU-1B	USG	BrEaST	BrEaST	0.63	0.62
		BrEaST	Ghana	0.38	0.53

Table 4. Performance metrics for primary and secondary test sets, where Mamm is Mammogram, USG is Ultrasound and ACC is accuracy.

4.1. Model Performance

The model was built and trained on ultrasound data from Ghana (Ghana) and BrEaST (Poland) and mammogram data from Ghana (Ghana), MIAS (UK), and INBreast (Portugal). The final model achieved a mean AUC of 0.8920, 0.5193, and 0.7226 when trained and tested with mammogram images from Ghana, MIAS, and INBreast respectively. The model achieved a mean AUC of 0.7415, and 0.6243 when trained and tested with ultrasound images data from Ghana and BrEaST, respectively. The final breast cancer classification model achieved an AUC of up to 0.8998 on ultrasound images and 0.8401 on mammogram images. These results were comparable with those of the state-of-the-art literature, where AUCs were found to range from 0.77 [2] to 0.92 [19].

In general, the model performed better on mammogram images than on ultrasound images. The model consistently performed highest on the images from Ghana. It is likely that this caused the higher number of images compared to the other publicly available datasets. The three public datasets did not perform as highly as the Ghana dataset. Between the two public mammogram datasets, the model trained and tested on the INBreast dataset performed higher than the model trained and tested on the MIAS dataset. Again, it is likely this was due to the INBreast dataset being larger with higher quality digital images.

4.2. Statistical Analysis

The objective of this work was to determine whether there is a change in performance when the model is tested with data

from the same demographic region as its training data, or a different region. The null hypothesis for the demographic experiments (shown in Table 5) was tested by applying the Student's t-test to the distribution of AUCs produced by the test sets from two demographics. The null hypothesis H_0 was rejected for Experiments DM-1A, DM-2A, DM-2B, DU-1A, and DU-1B with a p-value of 0.0000, indicating a significant change in performance when testing the model with a test set from a different demographic. The null hypothesis H_0 was only accepted for experiment DM-1B with a p-value of 0.2950 indicating that there was not a significant change in performance when the model trained with MIAS was tested with data from Ghana. It must be noted that the quality of the MIAS dataset was much lower than the other datasets.

ID	Prim. μ AUC	Sec. μ AUC	Statistic	p-value	Outcome
DM-1A	0.8290	0.5260	129.5625	0.0000	Reject
DM-1B	0.5193	0.5117	1.0548	0.2950	Accept
DM-2A	0.8279	0.5319	135.2443	0.0000	Reject
DM-2B	0.7225	0.5368	22.9138	0.0000	Reject
DU-1A	0.7415	0.5286	58.5347	0.0000	Reject
DU-1B	0.6243	0.5294	15.2902	0.0000	Reject

Table 5. Student's t-test outcome for H_0 of each experiment.

5. Conclusion and future Work

The presence of bias, regardless of the cause, has serious implications for the field of machine learning for breast cancer imaging. This study has demonstrated the risk of developing breast cancer classification models in one region and deploying them in another without careful consideration and bias mitigation. Further work should aim to isolate the possible causes of the bias that has been found. The effects of bias should also be investigated between different groups in a single dataset. Finally, research should show how such bias can be mitigated before the deployment of commercial AI for breast cancer detection.

Acknowledgment

Special thanks goes to George Asafu Adjaye Frimpong (radiologist), Spectra Health, Ghana, for providing us with radiological images.

References

- [1] Adewole S. Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*, 154(11):1247, 2018. 1, 2
- [2] Anton S. Becker, Magda Marcon, Soleen Ghafoor, Moritz C. Wurnig, Thomas Frauenfelder, and Andreas Boss. Deep learning in mammography diagnostic accuracy of a multi-purpose image analysis software in the detection of breast cancer. *Investigative Radiology*, 52(7):434–440, 2017. 5
- [3] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. 1
- [4] Alessandro Carriero, Léon Groenhoff, Elizaveta Vologina, Paola Basile, and Marco Albera. Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements in Early 2024. *Diagnostics*, 14(8):848, 2024. 1, 2
- [5] Karin Dembrower, Erik Wåhlin, Yue Liu, Mattie Salim, Kevin Smith, Peter Lindholm, Martin Eklund, and Fredrik Strand. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *The Lancet Digital Health*, 2(9):e468–e474, 2020. 2
- [6] Karin Dembrower, Alessio Crippa, Eugenia Colón, Martin Eklund, and Fredrik Strand. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*, 5(10):e703–e711, 2023. 2
- [7] Arman Ghavidel and Pilar Pazos. Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review. *Journal of Cancer Survivorship*, pages 1–25, 2023. 2
- [8] Thomas Glare. How to Use One-Vs-Rest and One-Vs-One for Multi-Class Classification, 2020. 4
- [9] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024. 3
- [10] Yang Gu, Wen Xu, and et. al. Deep learning based on ultrasound images assists breast lesion diagnosis in China: a multicenter diagnostic study. *Insights into Imaging*, 13(1): 1–14, 2022. 4
- [11] Jonathan Huang, Galal Galal, Mozziyar Etemadi, and Mahesh Vaidyanathan. Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review, 2022. 2
- [12] Charles Jones, Daniel C. Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence* 2024 6:2, 6(2):138–146, 2024. 2
- [13] Hyo Eun Kim, Hak Hee Kim, Boo Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun Kyung Kim. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2(3):e138–e148, 2020. 2, 3, 4
- [14] Learn Statistics Easily. What is the difference between t test vs mann whitney u test?, 2023. 5
- [15] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. INbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012. 3

- [16] Jung In Park, Selen Bozkurt, Jong Won Park, and Sunmin Lee. Evaluation of race/ethnicity-specific survival machine learning models for Hispanic and Black patients with breast cancer. *BMJ Health and Care Informatics*, 30(1), 2023. [1](#), [2](#)
- [17] Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data* 2024 11:1, 11(1):1–13, 2024. [3](#)
- [18] N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. von Karsa. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition - Summary document, 2008. [1](#)
- [19] Thomas et. al. Schaffter. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Network Open*, 3(3):E200265, 2020. [4](#), [5](#)
- [20] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B.A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine* 2021 27:12, 27(12):2176–2182, 2021. [2](#)
- [21] James Suckling. The mini-MIAS database of mammograms, 1994. [3](#)
- [22] Y Weerakkody, T Manning, P Lemos, and et al. Breast imaging-reporting and data system (BI-RADS) — Radiology Reference Article — Radiopaedia.org. [3](#)
- [23] Brennan Whitfield. An Introduction to the Shapiro-Wilk Test for Normality — Built In. [5](#)
- [24] Jenny Yang, Andrew A.S. Soltan, and David A. Clifton. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *npj Digital Medicine* 2022 5:1, 5(1):1–8, 2022. [2](#), [3](#)