SpurBreast: A Curated Dataset for Investigating Spurious Correlations in Real-world Breast MRI Classification

Jong Bum Won[1], Wesley De Neve[1,2], Joris Vankerschaver[1,3], and Utku Ozbulak[1,2] (✉)

[1] Center for Biosystems and Biotech Data Science, Ghent University Global Campus, Incheon, Republic of Korea
[2] IDLab, ELIS, Ghent University, Ghent, Belgium
[3] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
(✉) utku.ozbulak@ghent.ac.kr

Abstract. Deep neural networks (DNNs) have demonstrated remarkable success in medical imaging, yet their real-world deployment remains challenging due to spurious correlations, where models can learn non-clinical features instead of meaningful medical patterns. Existing medical imaging datasets are not designed to systematically study this issue, largely due to restrictive licensing and limited supplementary patient data. To address this gap, we introduce SpurBreast, a curated breast MRI dataset that intentionally incorporates spurious correlations to evaluate their impact on model performance. Analyzing over 100 features involving patient, device, and imaging protocol, we identify two dominant spurious signals: magnetic field strength (a global feature influencing the entire image) and image orientation (a local feature affecting spatial alignment). Through controlled dataset splits, we demonstrate that DNNs can exploit these non-clinical signals, achieving high validation accuracy while failing to generalize to unbiased test data. Alongside these two datasets containing spurious correlations, we also provide benchmark datasets without spurious correlations, allowing researchers to systematically investigate clinically relevant and irrelevant features, uncertainty estimation, adversarial robustness, and generalization strategies. Models and datasets are available at github.com/utkuozbulak/spurbreast.

Keywords: Spurious correlations · Breast cancer · Medical imaging.

1 Introduction

Deep Neural Networks (DNNs) have achieved remarkable success in medical imaging, demonstrating the potential to match or even surpass expert performance in diagnosing diseases [20,28]. Despite these advancements, their deployment in clinical settings remains challenging due to their susceptibility to distribution shifts – where differences between training and real-world data lead

to significant drops in performance [22,27]. A critical factor contributing to this issue is the presence of spurious correlations, which occur when models inadvertently learn associations between irrelevant features and target labels, instead of focusing on clinically meaningful patterns [8]. In medical imaging, such correlations can arise from demographic biases, scanner artifacts, or variations in clinical settings, leading to models that fail to generalize effectively and, in turn, pose substantial risks in real-world applications [14]. These unintended dependencies not only reduce diagnostic accuracy but also have broader implications, such as influencing clinical decision-making processes and potentially introducing biases in healthcare access and insurance claims.

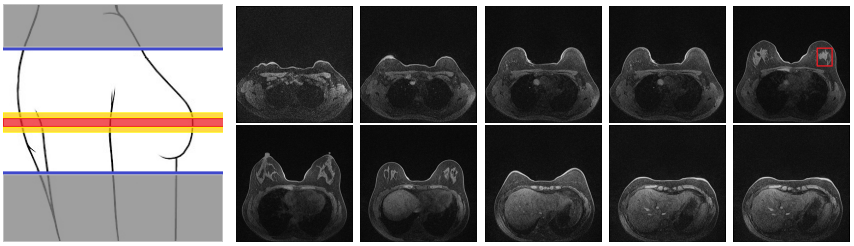
Although widely-used medical imaging datasets such as CheXpert, MURA, and MIMIC-CXR have facilitated the development of AI models [9,21,10], they are not specifically designed to investigate spurious correlations. Existing datasets tailored for this purpose, such as ImageNet-C/P and Spawrious [7,18], primarily focus on natural and synthetic images, which fail to capture the unique complexities of medical imaging.

Creating curated datasets in medical imaging containing well-documented spurious correlations presents unique problems, primarily due to licensing restrictions and regulatory guidelines that necessitate the use of de-identified and unaltered medical images [2,11]. While synthetic datasets have been proposed to circumvent these issues, they often lack realism and fail to capture the variability inherent in real-world medical imaging [25]. On the other hand, discovering naturally occurring spurious correlations is particularly difficult because they often stem from subtle and indirect relationships [19]. Furthermore, the presence of domain-specific biases and imbalanced data distributions exacerbates the problem, as certain demographic groups or disease types may be overrepresented, making it challenging to disentangle spurious associations from meaningful clinical features [16,26].

In this work, we introduce **SpurBreast**, a curated dataset designed to study spurious correlations in real-world breast MRI data. It consists of real-world patient data, carefully curated to include well-documented spurious correlations such as those related to patient demographics and imaging equipment. In addition, we have conducted an extensive experimental analysis on supplementary patient data to systematically assess the impact of spurious correlations on model performance. Unlike existing datasets, SpurBreast provides a comprehensive framework that allows researchers to study the influence of spurious correlations under controlled conditions, facilitating the development of more robust and generalizable AI models for medical imaging.

2 Data

MRI images. The data used in this study are from the DUKE Breast Cancer Dataset [23], a comprehensive single-institutional retrospective collection of 3D MRI scans from over 900 patients with biopsy-confirmed invasive breast cancer at a university hospital. Each study includes a 3D MRI acquired using 1.5T or 3T



(a) Side profile (b) Example MRI slices

Fig. 1: (a) A side profile diagram of the breast, highlighting the imaging region. Slices in the red area contain MRI images with breast tumors, slices in the yellow area are buffer zones and are not used, and slices in the white region do not contain invasive breast tumors. (b) Example MRI slices obtained from the specified cross-sectional region. Image with the highlighted red box in one slice indicates an invasive breast tumor.

scanners, from patients in the prone position. On average, each 3D scan consist of 250 2D slices (see Figure 1). For the predictive tasks, the slices are categorized into two groups: those containing breast tumors and those without. Following the approach of [15,12], we establish a buffer zone between slices containing tumors and those that do not (highlighted in yellow in Figure 1a). Images within this buffer zone are excluded from analysis, and the remaining slices are labeled and used for the predictive task.

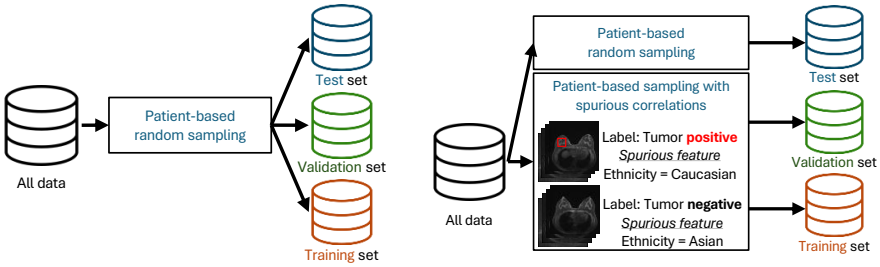
Supplementary information. Alongside the image data, separate tabular data cover various types of patient information including demographic, clinical, pathology, treatment information gathered from clinical notes, radiology reports, and pathology reports. Apart from that, these tabular data also encompass details about imaging devices and characteristics, including size, shape, texture, and enhancement patterns of both the tumor and surrounding tissue. Overall, these data contain more than 100 features. Unfortunately, the majority of these features are imbalanced in distribution, making the process of dataset creation challenging when taking those features into account.

3 Methodology

3.1 Discovering Spurious Correlations

In typical machine learning studies, the data are split into training, testing, and validation subsets randomly in order to prevent spurious correlations from arising (see Figure 2a). Then, the training set is used to optimize the model parameters, the validation set to tune hyperparameters and guide model selection, and the testing set to serve as a final evaluation metric for generalization on new data.

Different than the aforementioned approach, in this study, we aim to find features that spuriously correlate with predictive labels and create a well-documented



(a) Random split (b) Split with spurious correlations

Fig. 2: Illustration of the dataset creation process for discovering spurious correlations. (a) A typical patient-based random sampling approach, where the dataset is split into training, validation, and test sets to prevent overlap and ensure unbiased evaluations. (b) A modified sampling strategy where specific spurious correlations between predictive labels (tumor-positive and tumor-negative) and supplementary features (e.g., ethnicity) are deliberately introduced to study their effects on model performance.

dataset containing clearly defined spurious correlations. To achieve this, we adopt the approach detailed below.

Testing dataset. Before creating any training and validation datasets, we randomly select 150 patients and use the tumor-positive and tumor-negative slices from those images for the testing dataset. Throughout the manuscript, we use this dataset to consistently measure the generalization performance of trained models on unseen data and ensure that measurements on the test data are easily comparable.

Training and validation datasets. Using the supplementary features described in Section 2, we divide the training and validation data, at the patient level, in a such way that the positive labels in both datasets are associated exclusively with images possessing a specific property, while the negative labels are linked to images with a different property. For example, using the ethnicity feature, we select all tumor-positive images from Caucasian patients, whereas all tumor-negative images are selected from Asian patients (see Fig. 2b). This setup introduces a spurious correlation between ethnicity and the predictive label. If the spurious correlation is strong enough, models may exploit it to achieve high training and validation performance. However, their performance on the testing dataset, evaluated on data without spurious correlations, will be significantly lower.

3.2 Datasets and Evaluated Features

Using the approach outlined above, we create datasets based on a variety of unique features in the supplementary data. However, a substantial challenge in this process is data imbalance. For instance, only a small subset of patients in

our dataset exhibit nipple retraction. Consequently, based on this feature, patients cannot be effectively split into training and validation sets to train DNNs, as the sample size is insufficient for meaningful training. Another challenge is missing data; a large number of features have a high proportion of missing labels, rendering them impractical for use in dataset splits. Based on the available data, we investigate more than 100 features but only report on several interesting and relevant features due to space constraints:

Ethnicity. This feature describes the self-reported ethnic background of patients, such as Caucasian, Asian, or African American. Differences in representation across groups can introduce spurious correlations if ethnicity disproportionately aligns with specific labels.

Menopause status. This feature indicates a patient's menopausal status (e.g., premenopausal or postmenopausal), which affects breast tissue density. Its link to age and demographics may unintentionally correlate with diagnostic labels.

Magnetic field strength. The strength of the MRI scanner's magnetic field, measured in Tesla (e.g., 1.5T or 3T). Differences in field strength can affect image quality and introduce unintended correlations with labels.

Surgery type. This feature describes the type of surgery that will be performed on the patient (e.g., mastectomy or lumpectomy). Variability in surgical decisions may correlate with disease characteristics, leading to potential biases.

Vertical alignment. Indicates whether an image was vertically flipped during data augmentation. Uneven application of this transformation can create unintended associations with specific labels.

Baseline performance. Apart from the dataset splits created based on the aforementioned features, we also use three baseline performance datasets to evaluate the models on randomly sampled datasets without spurious correlations. These datasets are sampled based on patient counts and represent low-data, medium-data, and large-data benchmarks.

Details about these datasets, the number of images in the training, validation, and testing datasets, as well as the number of patients containing information regarding these features, are provided in Table 1.

3.3 Models

In the upcoming experiments, we use two models: a robust and widely adopted convolutional model in the literature (ResNet-50) [6], and a transformer-based architecture, Vision Transformer (ViT-B/16) [4], which has demonstrated state-of-the-art performance in various computer vision tasks. All models are initialized with pretrained weights from the ImageNet dataset [3], ensuring a strong starting point for transfer learning. The pretrained weights for these models are taken from the PyTorch library.

Table 1: The table presents the number of patients and images allocated to training, validation, and testing sets for different dataset configurations. The baseline datasets (low, medium, and large) contain randomly sampled patient data without spurious correlations, while the experimental datasets introduce specific spurious correlations based on features such as ethnicity, menopause status, MRI field strength, surgery type, and image alignments. The patient counts and corresponding image counts are shown for each dataset category.

Feature	Patients in dataset			Images in dataset		
	Training	Validation	Testing	Training	Validation	Testing
Baseline (Low-data)	150	150		6,490	8,032	
Baseline (Medium-data)	400	150		19,252	8,032	
Baseline (Large-data)	600	150		29,196	8,032	
Ethnicity	200	100	150	4,974	2,408	6,788
Menopause	400	150		10,488	4,330	
Field strength	400	150		9,562	3,576	
Surgery type	400	150		7,860	2,916	
Vertical alignment	500	150		24,138	6,926	

4 Experimental Results

We train the models described in Section 3.3 using the datasets provided in Table 1. To identify the best-performing model for each dataset, we conduct a comprehensive grid search and train the models for 50 epochs using three optimization algorithms: Stochastic Gradient Descent (SGD), Adam [13], and AdamW [17]. The initial learning rates are set to $10^{\{-5,-4,-3,-2\}}$. For SGD, we adopt a cosine annealing learning rate schedule, following the approach of [1]. Additionally, we experiment with weight decays of 10^{-4}, 10^{-5}, and 0. All models are trained with a batch size of 32, and the models achieving the highest accuracy on the validation set are selected as the final models.

The experimental results, summarized in Table 2 provide a comprehensive evaluation of model performance across different dataset configurations. We report accuracy (Acc.), positive predictive value (PPV), and negative predictive value (NPV) for training, validation, and testing datasets. The baseline datasets without spurious correlations serve as a reference point, while the experimental datasets, which include spurious features, allow us to assess the impact of spurious correlations on model generalization.

Consistent results with baseline datasets. The baseline datasets show minimal performance degradation from validation to testing, with stable accuracy and predictive values across different dataset sizes. This consistency suggests that DNNs generalize well to unseen data when no spurious correlations are present, reinforcing the dataset's validity.

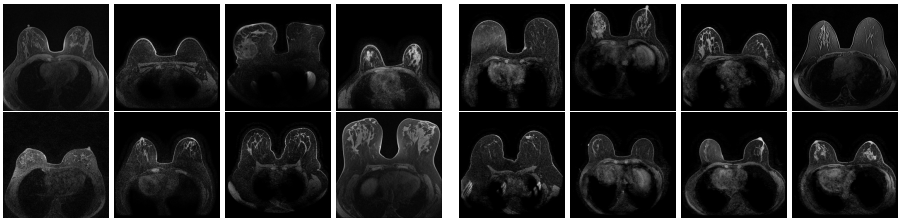
Weak or no spurious correlations. While models trained on datasets containing spurious correlations on ethnicity, menopause status, and surgery type show some decline in testing accuracy, the impact is not severe. The test accuracy for these features remains above 70%, indicating that although spurious correlations influence model predictions, their effect is not as dominant.

Table 2: Accuracy (Acc.), positive predictive value (PPV), and negative predictive value (NPV) for ResNet-50 and Vit-B models across training, validation, and testing datasets are provided, considering various experimental conditions, including randomized baseline data, demographic factors, and data augmentation techniques.

		Training			Validation			Testing		
Feature	Model	Acc.	PPV	NPV	Acc.	PPV	NPV	Acc.	PPV	NPV
Baseline (Low-data)	ResNet-50	0.77	0.69	0.84	0.71	0.76	0.66	0.77	0.78	0.75
	Vit-B	0.79	0.74	0.85	0.75	0.70	0.80	0.79	0.70	0.88
Baseline (Medium-data)	ResNet-50	0.76	0.66	0.86	0.75	0.77	0.73	0.77	0.77	0.77
	Vit-B	0.79	0.72	0.86	0.76	0.67	0.86	0.81	0.69	0.94
Baseline (Large-data)	ResNet-50	0.80	0.74	0.86	0.76	0.76	0.77	0.82	0.78	0.86
	Vit-B	0.81	0.75	0.87	0.79	0.71	0.87	0.83	0.72	0.94
Ethnicity	ResNet-50	0.97	0.96	0.97	0.85	0.89	0.81	0.72	0.75	0.68
	Vit-B	0.83	0.81	0.86	0.81	0.77	0.85	0.71	0.65	0.77
Magnetic Field Strength	ResNet-50	**0.99**	0.98	0.99	**0.99**	1.00	0.98	**0.52**	0.62	0.41
	Vit-B	**0.98**	0.98	0.98	**0.99**	0.99	0.99	**0.55**	0.66	0.43
Menopause	ResNet-50	0.91	0.91	0.92	0.85	0.88	0.82	0.71	0.77	0.65
	Vit-B	0.67	0.63	0.72	0.71	0.62	0.80	0.69	0.62	0.75
Surgery Type	ResNet-50	0.53	0.49	0.57	0.70	0.79	0.61	0.74	0.80	0.68
	Vit-B	0.80	0.73	0.87	0.75	0.63	0.87	0.77	0.66	0.88
Vertical Alignment	ResNet-50	**0.99**	0.99	0.99	**1.00**	1.00	1.00	**0.52**	0.04	1.00
	Vit-B	**0.98**	0.98	0.98	**1.00**	1.00	1.00	**0.52**	0.05	1.00

Strong spurious correlations. Models trained on datasets where magnetic field strength and vertical alignment are spuriously correlated with the target labels achieve near-perfect accuracy during training and validation but suffer a substantial drop in test performance, with accuracy declining to around 50%. This sharp decrease indicates that models are heavily relying on these non-clinical attributes for decision-making rather than learning meaningful medical features. In the case of magnetic field strength, the model labels all 1.5T images as tumor-positive and all 3T images as tumor-negative, while for vertical alignment, images facing up are predicted as tumor-positive and those facing down as tumor-negative. This confirms that the model learns to exploit these non-clinical cues as shortcuts, failing to generalize when tested on unbiased data.

Confirming strong spurious correlations. Based on the initial set of experiments provided above, we identify magnetic field strength and vertical alignment as the two features that, when spuriously correlated with image labels, lead models to learn these spurious signals instead of clinically meaningful features. To confirm these observations and ensure that the results presented in Table 2 are not merely one-off outcomes based on dataset sampling, we repeat the same experiment 10 times with different randomized patients in the training, validation, and testing datasets. In all of those experiments, we find that validation accuracy reaches $\sim 100\%$ at first few epochs while test accuracy remains $\sim 50\%$.



(a) MRIs from 1.5 Tesla device (b) MRIs from 3 Tesla device

Fig. 3: Example breast MRI images obtained using (a) 1.5T and (b) 3T devices.

4.1 Understanding Spurious Correlations

Magnetic field strength. 3T scanners offer higher magnetic field strength, improving signal-to-noise ratio (SNR) and image resolution for sharper, more detailed images [24]. However, they are more prone to artifacts, heating effects, and signal loss, especially around metal implants [5]. As such, our proposed dataset involving this spurious signal features a non-local spurious signal that influences the entire image rather than a localized region. An example set of images obtained from 1.5T and 3T devices are provided in Figure 3, showing that it is visually not possible to distinguish 1.5T MRIs from the 3T ones.

Vertical orientation. Different from magnetic field strength, which affects the entire image globally, vertical orientation is a local feature that only alters the spatial arrangement of structures within the image. This transformation does not modify the underlying tissue characteristics or signal properties but instead introduces artificial correlations that models may exploit as shortcuts.

5 Conclusions and Future Perspectives

We introduce **SpurBreast**, a curated dataset designed to study the impact of spurious correlations in breast MRI classification. It includes two experimental datasets with specific biases to evaluate model robustness. The first dataset introduces a spurious correlation with MRI magnetic field strength and the second dataset introduces spurious correlations based on vertical alignment. In addition to these datasets with spurious correlations, we provide a baseline dataset free of spurious correlations, serving as a benchmark for unbiased evaluation and bias mitigation.

Our goal in providing datasets with spurious correlations is to enable researchers to investigate how models learn and rely on unintended features, measure uncertainty, and develop methods to improve model generalization. Models and datasets are available at github.com/utkuozbulak/spurbreast.

Acknowledgments. The authors have no acknowledgments to declare.

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

References

1. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
2. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The cancer imaging archive (TCIA): Maintaining and operating a public information repository. Journal of Digital Imaging **26**, 1045–1057 (2013)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
5. Graves, M.J.: 3 t: the good, the bad and the ugly. British Journal of Radiology **95** (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
7. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
8. Hermann, K., Mobahi, H., FEL, T., Mozer, M.C.: On the foundations of shortcut learning. In: International Conference on Learning Representations (2024)
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19, AAAI Press (2019)
10. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6** (2019)
11. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3** (2016)
12. Kang, S., De Neve, W., Rameau, F., Ozbulak, U.: Exploring patient data requirements in training effective ai models for mri-based breast cancer classification. In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care. pp. 75–84. Springer (2024)
13. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
14. Koçak, B., Ponsiglione, A., Stanzione, A., Bluethgen, C., Santinha, J., Ugga, L., Huisman, M., Klontzas, M.E., Cannella, R., Cuocolo, R.: Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagnostic and Interventional Radiology (2024)

15. Konz, N., Gu, H., Dong, H., Mazurowski, M.A.: The Intrinsic Manifolds of Radiological Images and Their Role in Deep Learning, p. 684–694. Springer Nature Switzerland (2022)
16. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
18. Lynch, A., Dovonon, G.J., Kaddour, J., Silva, R.: Spawrious: A benchmark for fine control of spurious correlation biases. arXiv preprint arXiv:2303.05470 (2023)
19. Olesen, V., Weng, N., Feragen, A., Petersen, E.: Slicing through bias: Explaining performance gaps in medical image analysis using slice discovery methods. In: MICCAI Workshop on Fairness of AI in Medical Imaging. pp. 3–13. Springer (2024)
20. Pham, T.C., Luong, C.M., Hoang, V.D., Doucet, A.: AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-cnn architecture with custom mini-batch logic and loss function. Scientific Reports **11** (2021)
21. Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. In: Medical Imaging with Deep Learning (2018)
22. Saab, K., Hooper, S., Chen, M., Zhang, M., Rubin, D., Re, C.: Reducing reliance on spurious features in medical image classification with spatial specificity. In: Proceedings of the 7th Machine Learning for Healthcare Conference. vol. 182, pp. 760–784. PMLR (2022)
23. Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. British journal of cancer **119**(4), 508–516 (2018)
24. Schmitt, F., Grosu, D., Mohr, C., Purdy, D., Salem, K., Scott, K.T., Stoeckel, B.: 3 tesla MRI: successful results with higher field strengths. Radiologe **44**(1), 31–47 (2004)
25. Stanley, E.A., Souza, R., Wilms, M., Forkert, N.D.: Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data. EBioMedicine **111** (2025)
26. Vaidya, A., Chen, R.J., Williamson, D.F., Song, A.H., Jaume, G., Yang, Y., Hartvigsen, T., Dyer, E.C., Lu, M.Y., Lipkova, J., et al.: Demographic bias in misdiagnosis by computational pathology models. Nature Medicine **30**(4), 1174–1190 (2024)
27. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P.N., Thadaney-Israni, S., Goldenberg, A.: Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine **25**, 1337–1340 (2019)
28. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L.L.Y., Ho, K., Weinstein, J.D., Reig, B., Gao, Y., Toth, H., Pysarenko, K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S.G., Heacock, L., Moy, L., Cho, K., Geras, K.J.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Transactions on Medical Imaging **39**(4), 1184–1194 (2020)