Analysis of Potential Biases on Mammography Datasets for Deep Learning Model Development

Blanca Zufiria[1,2](✉), Karen López-Linares[1,3], María J. García[1],
Kristin M. Rebescher[1], Iván Lalaguna[4], Esther Albertín[4],
Maria B. Cimadevila[5], Javier Garcia[5], Maria J. Ledesma-Carbayo[2],
and Iván Macía[1,3]

[1] Vicomtech, Basque Research and Technology Alliance, San Sebastián, Spain
[2] Universidad Politécnica de Madrid, Madrid, Spain
blanca.zufiria@gmail.com
[3] Biodonostia Health Research Institute, San Sebastián, Spain
[4] Instrumentación y Componentes SA, Inycom, Zaragoza, Spain
[5] Servicio Gallego de Salud, Galicia, Spain

Abstract. The development of democratized, generalizable deep learning applications for health care systems is challenging as potential biases could easily emerge. This paper provides an overview of the potential biases that appear in image analysis datasets that affect the development and performance of artificial intelligence algorithms. Especially, an exhaustive analysis of mammography data has been carried out at the patient, image and source of origin levels. Furthermore, we summarize some techniques to alleviate these biases for the development of fair deep learning models. We present a learning task to classify negative and positive screening mammographies and analyze the influence of biases in the performance of the algorithm.

Keywords: Bias · Deep learning · Mammography · Breast cancer

1 Introduction

Recent advances in artificial intelligence (AI) in the medical field enable transforming large sets of images together with their annotations into predictive models using deep learning techniques. Such a model is expected to behave in an unbiased way to produce fair, objective decisions, without basing them on spurious attributes. However, AI algorithms can be biased towards certain input patterns, deriving unfair decisions dependent on the domain and not on the task to be solved. Biases may come from several origins, among which data-related biases frequently appear [1,2]. Thus, to prevent from a biased behavior and ensure a good generalization of deep learning models in real-world environments, special care must be taken during the creation of training datasets and the design and development of the models [3,4]. There are recent studies in the literature that

S. Wu et al. (Eds.): AMAI 2022, LNCS 13540, pp. 59–67, 2022.
https://doi.org/10.1007/978-3-031-17721-7_7

analyze bias in deep learning algorithms applied to medical images [2,3,5,6]. [7,8] perform an analysis of the impact of bias related to sociological factors such as sex, age, race or type of health insurance. [9] describe a methodology to clinically evaluate AI technology on medical images. [10] found a source of bias in patient age, which they mitigated with adversarial training. Similarly, [11] apply a multi-task strategy together with an adversarial training scheme to simultaneously detect and mitigate bias (sex and skin tone) in a skin lesion detection scenario. [8,12–15] analyze selection biases in chest X-ray datasets and [8,14,15] emphasize on how acquisition equipment-related biases and domain shifts affect a pneumonia detection algorithm. Regarding mammography solutions, [2] comments that the presence of an image marker could interfere in the performance of the algorithm. [16,17] develop a deep learning algorithm to predict breast cancer risk and they use adversarial training to discriminate image origin, even if the variability in the manufacturers used during training is scarce. Furthermore, [18] developed an screening algorithm to predict cancer probability from a mammogram view using a wide variety of manufacturers. Nevertheless, they do not mention preprocessing techniques or data cleaning, which could derive into biases.

This paper aims at highlighting the relevance of performing an analysis of potential data-related biases before deep learning model development. Here, data bias is defined as gathered data that does not represent the phenomenon to predict. It can also contain characteristics produced by humans that may lead algorithms to solve a different task from the desired one and to fail when tested on properly selected independent data. In Sect. 2, we provide an overview on bias detection and mitigation techniques using a mammography dataset, with a high variability in manufacturers and models, as an example. Also, we show the influence of data related bias on classification experiments, together with possible solutions to reduce the impact of the bias. In Sect. 3, results from experiments are discussed. Finally, conclusions are provided in Sect. 4.

2 Materials and Methods

This section describes the input mammography dataset and our approach to analyze biases. We also present some techniques that can be used to mitigate these biases. Finally, some experiments were carried out to evaluate the influence of biases in deep learning algorithms.

2.1 Mammography Dataset

The dataset is composed of 1727 mammography studies provided by the Galician health care system. Since the goal was to provide democratized deep learning solutions for the health area, the main criterion to gather the data from the picture archiving and communication system (PACS) was to contemplate all the available manufacturers. Mammograms from Fujifilm Corporation, Hologic Inc, Philips Medical Systems, and Siemens were obtained and filtered so that only those containing two views, i.e. bilateral craniocaudal (CC) and mediolateral oblique (MLO), for each breast were considered. Finally, a selection according to the following breast cancer screening clinical categories was performed: 1) **Negative screening**: mammograms where radiologist did not detect signs of cancer and from women that were not derived for further tests and 2) **Positive screening**: mammograms where radiologists detected a sign of cancer and women were derived to further tests in the diagnostic departments. The distribution of exams in these categories, shown in Table 1, was balanced for most equipment manufacturers except for Philips scans, which were not used to acquire the mammography prior to further tests in any case.

Table 1. Number of studies distributed by manufacturers and clinical categories.

	Fujifilm	Hologic	Philips	Siemens	Total
Negative screening	277	263	271	270	1081
Positive screening	262	197	0	187	646
Total	539	460	271	457	1727

2.2 Bias Analysis

An in-depth analysis of the dataset for AI model development is an important step to detect potential biases and to ensure model performance in real world applications. Especially, datasets containing medical images are ideally built gathering information from different hospitals, different devices and several protocols to fulfill the needs of the whole health care system. Socio-technological analysis is crucial in these cases to detect potential biases, some of which can be discussed at the DICOM metadata level or at the content or pixel data level.

DICOM Metadata Analysis: Some information about the patient and the imaging studies can be directly extracted from standard DICOM tags (Fig. 1-a). In general, there are relevant differences between negative and positive studies. Specially, images are acquired with different scanners (Device ID) and acquisition parameters (WW/WC) between negative and positive exams. Furthermore, differences in the Patient Age and Institution tags induce a very important bias in the dataset. Hence, suggesting that negative studies may come from hospitals where a breast cancer screening program is carried out, whereas positive exams may come from diagnostic departments. Thus, when designing algorithms, the global performance of the network could be unfairly biased towards some specific devices, which should be detected and considered.

Histogram Analysis: Understanding the distribution of image intensity values across different categories is another approach to measure bias in the dataset and to decide appropriate preprocessing methods. Mean and standard deviation histograms are calculated (Fig. 1-b) for positive and negative screening exams independently. Differences are observed, probably related to the different scanners and acquisition parameters previously discussed and shown in Fig. 1-a.

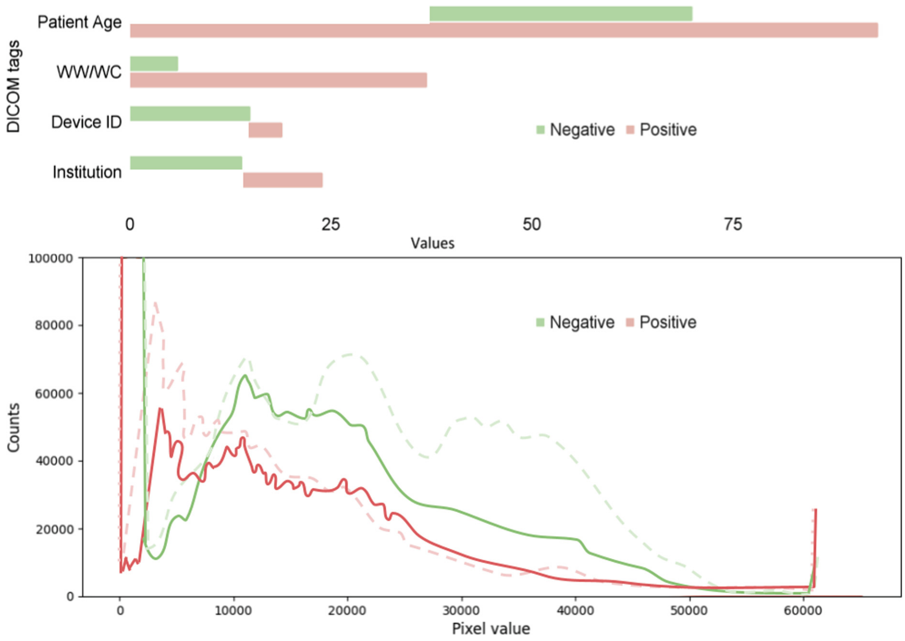

Fig. 1. a) Distribution positive and negative screening studies with respect to the different DICOM tags. b) Mean (continuous line) and standard deviation (dotted line) histograms for the different clinical categories in the dataset.

2.3 Bias Correction Techniques

Based on the analysis previously described, we identified some methods to mitigate the biases. The techniques can be divided into modification of image appearance or modification of model training and architecture to guide the learning towards the desired features.

Image Appearance Bias Correction: Some manufacturers introduce text marks in the image, e.g. labels indicating the view (CC, MLO) or the breast (left, right) and deletion of this markers is important to avoid biases. Furthermore, changing window width (WW) and window center (WC) values according to the VOI LUT DICOM tag of the study equalizes appearance between different manufacturers, devices and acquisition protocols.

Model Training Bias Correction: Domain-Adversarial training performs a domain transfer where final predictions must be made based on features that cannot discriminate the domain from which the images are obtained [19]. It could be a good solution to mitigate domain biases, derived from the distribution of the different mammography units and hospitals found in the dataset. Furthermore, data augmentation during training can be used to mitigate biases by increasing the number of training samples with different appearances applying, for example, random Gaussian noise, elastic deformations and modifying the contrast and the brightness.

2.4 Experimental Setup

To show the influence of data-related bias, we carried out a classification task using a deep learning approach. Especially, we aim at building a model that differentiates between normal mammograms (negative screening) and mammograms from patients derived to further tests in the diagnostic departments of the hospital (positive screening).

We employ a network architecture based on [18] using the DenseNet121 architecture where the four instances (right CC, right MLO, left CC, left MLO) are used to decide whether a study is a negative or positive screening exam. The model combines features between breast views and it is trained to minimize a binary cross-entropy loss, with a learning rate of $1e^{-5}$, batch size of 4 and Adam optimizer. The dataset is divided into training (70%), validation (20%) and test (10%) for each class to train the network (manufacturers and clinical categories are balanced between subsets). First, images are rescaled between 0 and 1 and normalized dividing each image by the mean and the standard deviation of the intensities, calculated beforehand for the whole rescaled dataset. Studies acquired with inverted gray scale values are modified so all images have a dark background. Instances corresponding to the left breast are flipped to the right side to facilite the learning process. Finally, the training dataset is balanced according to the clinical categories to avoid a bias towards the majority class. Several experiments are performed to evaluate the influence of the bias for the screening classification task:

Baseline: the neural network is trained with the preprocessed dataset and parameters as described above. The aim is the classification of mammography studies into positive and negative screening focusing on breast tissues.

WW/WC: from the baseline, this experiment aims at adjusting WW and WC values of the mammograms to homogenize the images across the acquisition devices (Fig. 3-f) as described in Sect. 2.4.

Data Augmentation: data augmentation is included to the WW/WC experiment as described in Sect. 2.4.

Domain-Adversarial Training: the goal is to obtain device independent features to mitigate the image type bias and focus more on the clinical classification task (Sect. 2.4). Based on the fact that a model could be trained to differentiate between devices (Fig. 3-e, upper figure), a domain-adversarial training that extracts intermediate features independent on the device could be developed. We introduce a domain-discriminator to classify features from different devices according to the Device ID DICOM tag and thereby, encourage similar feature extraction for all the domains to solve the actual screening classification task

(Fig. 3-e). The training procedure minimizes the loss of the classifier differentiating between negative and positive samples while maximizing the loss of the domain classifier.

Unbiased Data Addition: the inclusion of additional unbiased data could help the network focusing on the desired clinical task by ignoring previous biases. Hence, a new dataset was requested from screening units with balanced manufacturers and devices (Fig. 2) for 1179 screening negative and 393 screening positive mammograms. Furthermore, all these patients belong to the breast cancer screening program so the age range is fixed (Fig. 2). Preprocessing (with WW/WC modifications) of the images and data augmentation were applied in this experiment.

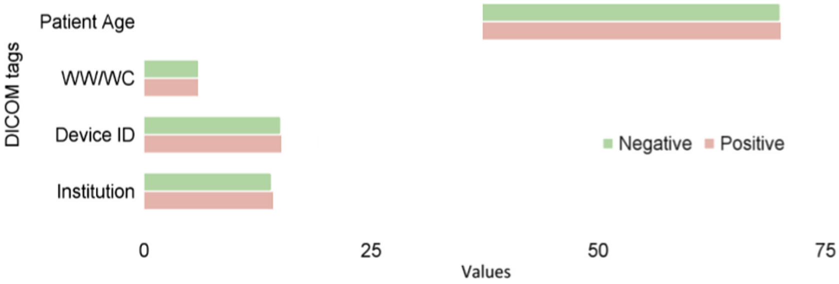

Fig. 2. Distribution positive and negative screening studies with respect to the different DICOM tags for the unbiased dataset. The distribution of the tags is balanced unlike in Fig. 1-a.

3 Results and Discussion

Trained models are evaluated on a subset of 188 mammography studies (104 negative screening and 84 positive screening) separated from the dataset and on a subset of 156 studies (125 negative screening and 31 positive screening) from the additional unbiased dataset. The experiments yield similar results, as shown in Table 2, where models achieve a high performance on test studies but metrics worsen for the unbiased test dataset, suggesting that the bias is not overcome (Table 2). Positive screening studies are misclassified as negative screening studies, probably influenced by its origin of acquisition (scanners and hospitals).

An extra verification of the models performance was carried out by visualizing their learning with heatmap explanations applying the Grad-CAM algorithm [20] (Fig. 3). Grad-CAM produces a coarse localization map that highlights the important regions in the image used to predict a specific class. Results shown in Fig. 3 suggest that screening models are not classifying studies according to the desired clinical task, as they focus more on the type of image than on breast

Table 2. Evaluation metrics on test and unbiased test subsets (separated before experiments training)

	Test dataset (188 exams)			Unbiased test dataset (156 exams)		
	ROC-AUC	*Sensitivity*	*Specificity*	*ROC-AUC*	*Sensitivity*	*Specificity*
Baseline	0.997	0.988	0.961	0.501	0.064	0.968
WW/WC	0.996	0.976	0.961	0.503	0.032	0.992
Data augmentation	0.982	0.988	0.923	0.525	0.064	0.944
Domain adversarial	0.995	0.952	0.971	0.551	0.032	1.0
Unbiased data addition	0.995	0.964	0.961	0.658	0.032	1.0

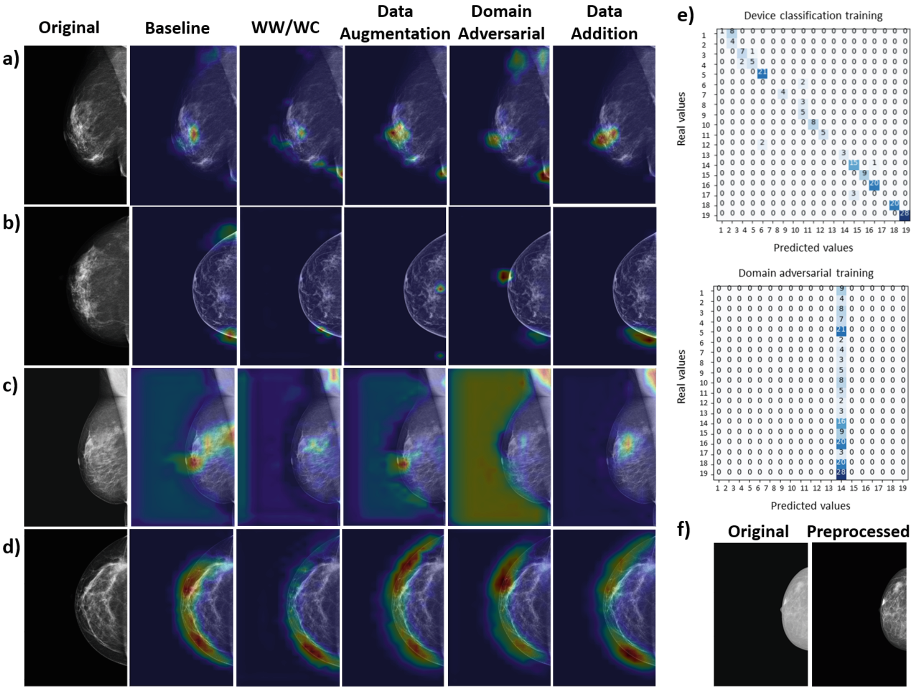

Fig. 3. (a–d) Grad-CAM computed for correctly classified mammography studies from the test subset. a) Right MLO view of a negative Siemens study. b) Left CC view of a negative Hologic stud. c) Left MLO view of a positive Fujifilm study. d) Left CC view of a positive Philips study. e) Confusion matrices of the device classifier (upper figure) and the domain discriminator classifier during the domain-adversarial training (lower figure). f) Preprocessed instance modifying the window width (WW) and window center (WC) values according to the function defined in the VOI LUT DICOM tag (Sect. 2.4).

tissues to find abnormalities. This is visible in Fig. 3-d, where a mass is present but models focus on the curvature of the breast. Furthermore, such explanations are highlighted on other parts of the images outside the breast like illuminated borders(Figure 3-b) or background (Fig. 3-c). Adding unbiased data mitigates the bias in some cases, where the trained algorithm focuses on the tissues inside the breast and not on the background or other characteristics derived from the device (Fig. 3-c). However, such mitigation is not enough to train a fair algorithm as seen in the quantitative results on the unbiased test subset (Table 2). Finally, domain-adversarial training results show that the discriminator is not able to differentiate between mammography devices (Fig. 3-e, upper figure) but metrics on the unbiased dataset (Table 2) and Grad-cam visualizations (Fig. 3) demonstrate that the bias still persists. Hence, based on the quantitative results and the Grad-CAM visualizations on test studies, we assume that models are biased not only by image-related features, as shown in the presented experiments, but also due to other the patient-related characteristics, such as age.

4 Conclusions

Hereby, we presented a bias analysis approach for deep learning applications that focuses on the inspection of DICOM metadata and pixel data distribution, using a mammography dataset as use case. Bias correction techniques were proposed and evaluated with experiments proving that, for the specific clinical task of breast cancer screening, results are biased toward the source of origin. Further techniques, like transfer learning, should be implemented to mitigate the existing biases in the mammography dataset. Such biases could be the age of patients, the acquisition techniques or other characteristics present in the two different screening and diagnostic departments. A careful initial inspection of the dataset before model building is essential to detect potential biases that may lead to unfair performance of AI algorithms. Hence, the proposed approach could help future researchers on the implementation of fair deep learning algorithms and methodology for dataset extraction and generation for medical imaging applications. Future work is needed to further investigate this potential issue with more experiments, analysis and techniques and to expand the research to different datasets.

Prospect of Application: assist in the development of fair AI models and unbiased database construction. Specially, in breast cancer screening scenario the robustness of AI models would increase for a fair performance at different health care systems, mammographers and acquisition protocols. Thus, all professionals and patients, regardless of the hospital they are in, would have access to the system equally.

Acknowledgments. This work has been partially funded by FEDER "Una manera de hacer Europa". This research has been done within the project CADIA - Sistema de Detecciónn de Diversas Patologíss Basado en el Analisis de Imagen con Inteligencia Artificial (DG-SER1-19-003) under the Codigo100 Public Procurement and Innovation Programme by the Galician Health Service - Servizo Galego de Saude (SERGAS) cofunded by the European Regional Development Fund (ERDF).

References

1. Hammer, G.P., du Prel, J.B., Blettner, M.: Avoiding bias in observational studies: part 8 in a series of articles on evaluation of scientific publications. Dtsch Arztebl Int. **106**, 664 (2009)
2. Yu, A.C., Eng, J.: One algorithm may not fit all: how selection bias affects machine learning performance. Radiographics **40**, 1932–1937 (2020)
3. Varoquaux, G., Cheplygina, V.: How I failed machine learning in medical imaging - shortcomings and recommendations. Electr. Eng. Syst. Sci. (2021)
4. Tong, S., Kagal, L.: Investigating bias in image classification using model explanations. Comput. Sci. Comput. Vis. Pattern Recogn. (2020)
5. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Re, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Comput. Sci. Mach. Learn. (2019)
6. K. Winkler, et al.: Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol (2019)
7. Pot, M., Kieusseyan, N., Prainsack, B.: Not all biases are bad: equitable and inequitable biases in machine learning and radiology. Insights Imaging **12**, 1–10 (2021)
8. 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. In: Proc. Natl. Acad. Sci. USA , **117**, 12592–12594 (2020)
9. Park, S.H., Han, K.: Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology **286**, 800–809 (2003)
10. Zhao, Q., Adeli, E., Pohl, K.M.: Training confounder-free deep learning models for medical applications. Nat. Commun. **11**, 1–9 (2020)
11. Li, X., Cui, Z., Wu, Y., Gu, L., Harada, T.: Stimating and improving fairness with adversarial learning. Comput. Sci. Comput. Vis. Pattern Recogn. (2021)
12. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. **27**, 2176–2182 (2021)
13. Catala, O.D.T., et al.: Bias analysis on public x-ray image datasets of pneumonia and Covid-19 patients. IEEE Access. **9**, 42370–42383 (2021)
14. E. H. P. Pooch, P. L. Ballester, R. C. Barros: Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. Electr. Eng. Syst. Sci. Image Video Process. (2020)
15. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. **15**, e1002683 (2018)
16. Yala, A., et al.: Toward robust mammography-based models for breast cancer risk. Sci Transl. Med. **13**, eaba4373 (2021)
17. Mayer McKinney, S., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**, 89–94 (2020)
18. Wu, N., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging **39**, 1184–1194 (2020)
19. Ganin, Y., et al.: Domain-adversarial training of neural networks. Stat. Mach. Learn. **17**, 2096–2130 (2016)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (ICCV). (2017)