Analysis of potential biases on mammography datasets for deep learning model development

Blanca Zufiria[1,2] BZUFIRIA@VICOMTECH.ORG
Karen López-Linares [1,3] KLOPEZ@VICOMTECH.ORG
María Jesús García [1] MJGARCIA@VICOMTECH.ORG
Kristin May Rebescher[1] KMREBESCHER@VICOMTECH.ORG
[1] *Vicomtech, Basque Research and Technology Alliance, Spain*
[2] *Universidad Politécnica de Madrid, Spain*
[3] *Biodonostia Health Research Institute, Spain*

Marcos Rubio[4] MARCOS.RUBIO@INYCOM.ES
Esther Albertín [4] ESTHER.ALBERTIN@INYCOM.ES
David Chaparro [4] DAVID.CHAPARRO@INYCOM.ES
[4] *Instrumentación y Componentes SA, Inycom, Spain*

Maria Blanca Cimadevila[5] MARIA.BLANCA.CIMADEVILA.ALVAREZ@SERGAS.ES
Javier Garcia[5] JAVIER.GARCIA.NAVAS@SERGAS.ES
[5] *Servicio Gallego de Salud, Spain*

Maria J. Ledesma-Carbayo [2] MLEDESMA@DIE.UPM.ES
Iván Macía[1,3] IMACIA@VICOMTECH.ORG

Editors: Under Review for MIDL 2022

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 computer-aided algorithms. Furthermore, we summarize some techniques to alleviate these biases. Particularly, we focus on possible biases on a mammography dataset and we present a classification task to 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 (GP et al., 2009; Yu and Eng, 2020). This type of bias can be due to class-imbalance or due to socio-technological factors. 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 (Varoquaux and Cheplygina, 2021; Tong and Kagal, 2020). Applying techniques to analyze and mitigate bias related to the data is an essential initial step in the development of deep learning models.

There are recent studies in the literature that analyze bias in deep learning algorithms applied to medical images. (Pot et al., 2021; Larrazabal et al., 2020b) perform an analysis of the impact of bias related to sociological factors such as sex, age, race or type of health insurance. (Park and Han, 2003) describe a methodology to clinically evaluate artificial intelligence technology on medical images, and they analyze, among other things, spectrum biases. (Zhao et al., 2020) explores how to fairly diagnose HIV from magnetic resonance images. They found a source of bias in patient age, which they mitigated by automatically extracting bias-free features from the image with adversarial training. Similarly, (Li et al., 2021) 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. (Seyyed-Kalantari et al., 2021; Catala et al., 2021; Larrazabal et al., 2020a) analyze selection biases in chest X-ray datasets. (Yu and Eng, 2020; Varoquaux and Cheplygina, 2021; Oakden-Rayner et al., 2019; Winkler et al., 2019) analyze potential biases and its importance on deep learning algorithms, where they focus on imaging-based problems and their additional factors to consider in their distribution: spectrum of imaging manifestations of disease and of normal appearances, imaging equipment, protocols used and medical intervention biases. (Pooch et al., 2020; Zech et al., 2018; Larrazabal et al., 2020b) demonstrate biases on chest X-ray datasets emphasizing how acquisition equipment-related biases and domain shifts affect a pneumonia detection algorithm. Regarding mammography solutions, (Yu and Eng, 2020) comments on a mammography dataset where the presence of an image marker could interfere in the performance of the algorithm. Finally, (Yala et al., 2021) develop a deep learning algorithm to predict breast cancer risk and they use adversarial training to discriminate image origin.

This paper aims at highlighting the relevance of performing an analysis of potential data-related biases before deep learning model development. In Section 2, we provide an overview on bias detection and mitigation techniques using a mammography dataset as an example. Also, we show the influence of data related bias on simple classification experiments, together with a possible solution to reduce the impact of the bias. In Section 3, results from experiments are discussed. Finally, conclusions are provided in Section 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.

2.1. Mammography dataset

The dataset for our analysis 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 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 4 clinical categories was performed:

- **Control**: mammograms where no abnormalities were detected by radiologists.
- **Benign**: mammograms where some benign lesion or cyst was detected by radiologists, but women were not derived for further tests. It was checked that no abnormalities were detected in two consecutive studies.
- **Biopsy-benign**: mammograms where a biopsy-proven benign lesion was found. The mammogram corresponds to the last exam taken before the biopsy.
- **Biopsy-malignant**: mammograms where a biopsy-proven malignant lesion was found. The mammogram corresponds to the last exam taken before the biopsy.

The distribution of exams in these categories, shown in Table 1, was balanced for most equipment manufacturers. However, Philips scans were not used to acquire the last mammography prior to biopsy in any case.

Table 1: Distribution of mammograms. Number of studies of the 4 clinical categories and 4 manufacturers

	Fujifilm	**Hologic**	**Philips**	**Siemens**	**Total**
Control	139	132	137	134	542
Benign	138	131	134	136	539
Biopsy-benign	134	100	0	63	297
Biopsy-malignant	128	97	0	124	349
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. Specially, 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 are crucial in these cases to detect potential biases, which can be discussed at the DICOM metadata level or at the content or pixel data level.

DICOM metadata analysis

Patient information and characteristics of the imaging studies can be directly acquired from standard DICOM tags. Among them, we focus our attention at three relevant groups: one relative to the device and general acquisition equipment, another group relative to the

presentation of the images and VOI LUT, and one group relative to patient information. We consider that these three groups are the most relevant to detect biases.

For the first group, the analyzed DICOM tags are Manufacturer, Manufacturer Model Name, Institution Name and Detector ID. The distribution of these tags in our dataset is summarized in Table 2. There are 4 manufacturers, 6 manufacturer models, 24 Institutions and 19 devices in total. We can see that manufacturers and models 1-3 (Fujifilm Corporation, Hologic, Inc., and Siemens) are shared for the four study categories, while manufacturer and model 4 (Philips Medical Systems) appears only for control and benign groups. Models 5 and 6 are only employed to acquire images in groups biopsy-benign and biopsy-malignant. On the other hand, the Institution and DeviceID are different between control/benign studies and biopsy studies. This huge difference induces a very important bias in the dataset, suggesting that control and benign studies may come from hospitals where a breast cancer screening program is carried out, whereas biopsy-benign and biopsy-malignant exams may come from diagnostic departments.

Table 2: Comparison of 4 different DICOM tags relative to the device for the 4 different clinical categories

	Manufacturer	**Model**	**Institution**	**Device ID**
Control	1-4	1-4	1-14	1-15
Benign	1-4	1-4	1-14	1-15
Biopsy-Benign	1-3	1-3, 5-6	15-24	16-19
Biopsy-Malignant	1-3	1-3, 5-6	15-22	16-18

For the second group of metadata, we analyzed the DICOM tags summarized in Table 3: Image Type, Window Center/Window Width, VOI LUT Function, Presentation LUT Shape, Presentation Intent Type. In this case, differences in tags between study categories are more subtle. However, there are relevant differences on Window Center and Window Width values between control/benign and biopsy studies, which affect the appearance of the image in terms of contrast and brightness. Furthermore, there are differences inside each category as images are acquired with different scanners and acquisition parameters (as described in the DICOM tags). 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 for each image type. The mean and standard deviation of the histogram of each subgroup is plotted in Figure 1, which shows differences between control/benign histograms versus biopsy-benign/biopsy-malignant histograms (related to the differences in the image aspects as shown in Table 3).

Table 3: Comparison of 5 different DICOM tags relative to image aspects for the different subgroups

Subgroup	**ImageType**	**VOILUT**	**LUTShape**	**IntentType**	**WW/WC**
Control	1-5	1-2	1-2	1	1-4/1-5
Benign	1-5	1-2	1-2	1	1-3,5/1-4,6
Biopsy-Benign	1-9	1	1-2	1	1,6-16/1-2, 7-17
Biopsy-Malignant	1-10	1	1-2	1	1,6-7, 17-36/1-2, 18-37

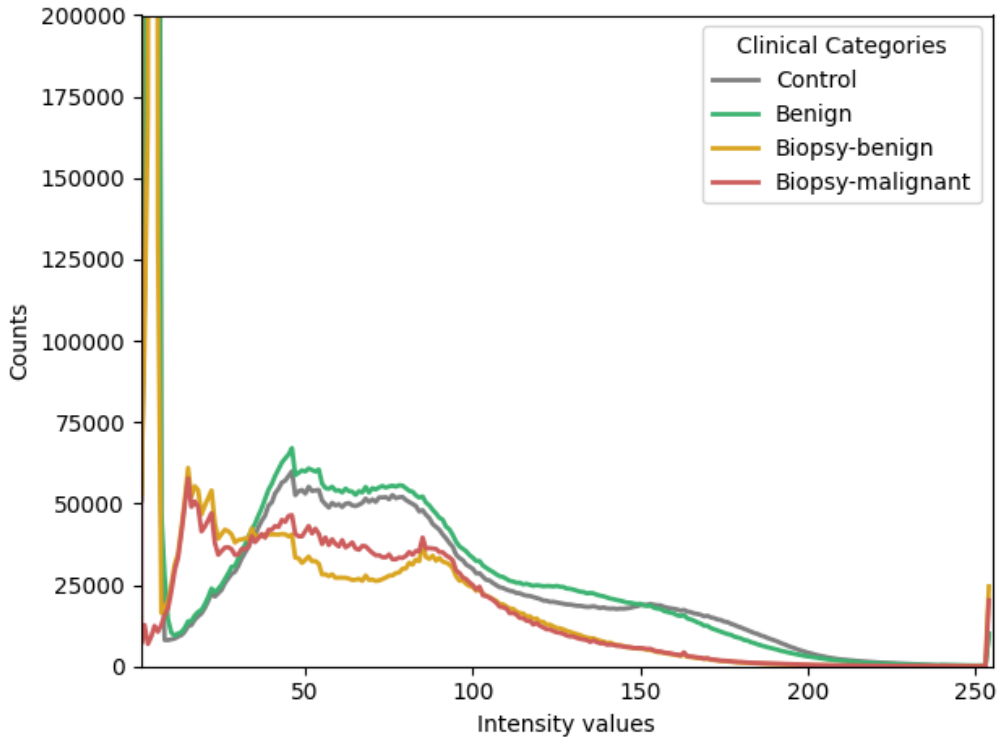

Figure 1: Mean histograms for the different study categories in the dataset.

Age analysis

A demographic analysis based on the age of the patients was performed to complement the two previous analysis. As shown in Figure 2, women from control and benign cases are between 48 and 71 years old (mean:55.78, stdv:40.59), while women that underwent a biopsy are between 31 and 93 years old (mean:31.19, stdv:30.36). This suggests, as also seen in Section 2.2, that women from control and benign studies may be inside a screening program,

whereas women that underwent a biopsy come from the diagnostic departments. This factor agrees with the ages of the women selected for the screening program at SERGAS.

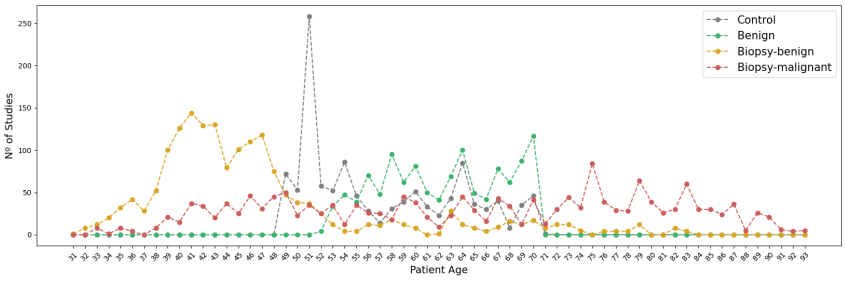

Figure 2: Ages of the patients for the different clinical groups

2.3. Bias correction techniques

Based on the bias analysis previously described, we identified the following methods to mitigate these biases: techniques that modify image appearance and techniques to modify the training approach and the model architecture to guide the learning towards the desired features.

Modification of image appearance

Images in the dataset can be preprocessed to modify their appearance by changing pixel data to mitigate biases. Some examples of these preprocessing techniques are:

- **Intensity windowing (IW)**: windowing with respect to the intensity histogram could be applied to images to equalize intensities between different manufacturers.
- **Histogram equalization (HE)**: distributes the intensities along the whole histogram to increase the contrast in the image.
- **Background label deletion**: some manufacturers introduce text marks in the image, e.g. labels indicating the view (CC, MLO) or the breast (left, right), which could introduce a bias when classifying studies.

Modification of model training

- **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 (Ganin et al., 2016). It could be a good solution to mitigate domain biases, derived from the distribution of the different mammography units and hospitals, found in the mammography dataset.

- **Data augmentation**: data augmentation can be used to mitigate biases by generating images with different intensities inside groups and subgroups of the dataset. This technique could be implemented by randomly applying different brightness and contrast in the images, increasing the number of training samples and generalizing their appearance (Appendix D-Equation (1)).

2.4. Experimental setup

To show the influence of data-related bias, we carry out some experiments to classify two different groups of exams.

We employ a network architecture based on (Nan Wu et al., 2020), where the four instances (right CC, right MLO, left CC, left MLO) of the study are used to decide whether it is a negative or positive screening exam. Specially, we trained CC and MLO networks (Densenet121) and then combined the features between breast views. We train our model for 30 epochs to minimize a binary cross-entropy loss, with a learning rate of 1^{-5}, batch size of 2 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. Data augmentation according to Appendix D-Equation (1) is applied and images are resized to 1024x512. Instances corresponding to the left breast are flipped to the right side to help the network focus on one side of the image, facilitating the learning task. Finally, the training dataset is balanced during training to avoid a bias towards the majority class.

- **Screening classification**: in this experiment, we aim at building a model to differentiate between normal mammograms and mammograms from patients with a derived biopsy, i.e, separate control and benign (negative) exams from biopsy-benign and biopsy-malignant (positive) exams.

- **Malignancy classification**: similar as the screening classification but the dataset is separated into benign cases (Control, Benign and Biopsy-benign) and malign cases (Biopsy-malign).

- **Screening domain-adversarial classification**: an adversarial training approach to obtain device-independent features could be interesting to mitigate the image type bias and focus more on the clinical classification task. First we aim at classifying images according to the mammography device, separating the data according to the "DeviceID" DICOM tag. The classification network successfully learned to identify studies according to the acquisition device. Based on the fact that the model learned to differentiate between devices, a domain-adversarial training that extracts intermediate features independent on the device is developed. The training procedure minimizes the loss of the label classifier (differentiating between negative and positive samples) while maximizing the loss of the domain classifier (according to the DeviceID). The results obtained showed that the discriminator was not able to differentiate between mammography devices.

3. RESULTS AND DISCUSSION

To visualize the explanations of the learning, we applied the Grad-cam approach (Selvaraju et al., 2017) that propagates the gradients of a predicted class backwards into the final convolutional layer. It produces a coarse localization map that highlights the important regions in the image used to predict a specific class. Grad-cam results suggest that the screening and malignant models are not classifying studies according to the desired clinical task. As shown in Appendix - Figure 4 and Figure 5, the model focuses more on the type of image (such as the background labels specific of the device) than on breast tissues to find abnormalities. Finally, a prospective test set was extracted form the screening department of the healthcare system (Appendix - Table 5) to evaluate the classification models. Results confirmed that the algorithm is biased towards the acquisition department origin (Appendix - Figure 6).

Table 4: Evaluation metrics on the test dataset for the different classification experiments.

EXPERIMENT	**ROC AUC**	**Error Rate**	**Sensitivity**	**Specificity**
Screening	0.996	0.0279	0.985	0.964
Malignancy	0.762	0.212	0.411	0.901
Screening domain-adversarial	0.937	0.062	0.995	0.901

Based on the Grad-cam visualizations and the validation tests, we confirmed that the classification clinical tasks are biased by the data distribution. Futhermore, a feature extractor independent on the acquisition device was built and could be used to mitigate this specific device bias during the learning. However, further techniques should be implemented to mitigate other existing biases in the mammography dataset such as the age or the acquisition techniques present in the two different screening and diagnostic departments.

4. CONCLUSIONS

Hereby, we presented a bias analysis approach for deep learning applications that focuses on the inspection of DICOM metadata, pixel data and age distribution, using a mammography dataset as use case. Significant differences are observed from this analysis between images acquired with different acquisition parameters and scanners, that could lead to unfair performances on deep learning methods. Bias correction techniques were proposed for mammography datasets based on the modification of the pixel data or the model training. Finally, some experiments were performed where we proved that for a specific clinical task, the results are biased toward the scanners and type of images and that a feature extractor invariant of the device could be trained to help mitigate this specific bias. 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.

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

Omar Del Tejo Catala, Ismael Salvador Igual, Francisco Javier Perez-Benito, David Millan Escriva, Vicent Ortiz Castello, Rafael Llobet, and Juan-Carlos Perez-Cortes. Bias analysis on public x-ray image datasets of pneumonia and covid-19 patients, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016.

Hammer GP, du Prel JB, and Blettner M. Avoiding bias in observational studies: part 8 in a series of articles on evaluation of scientific publications, 2009.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, and Enzo F et al Diego H. Milone. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, 2020a.

J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, 2020b.

Xiaoxiao Li, Ziteng Cui, Yifan Wu, Lin Gu, and Tatsuya Harada. Estimating and improving fairness with adversarial learning, 2021.

Jason Phang Nan Wu, Jungkyu Park, Yiqiu Shen, Masha Zorin Zhe Huang, and Stanislaw Jastrzebski et al. Deep neural networks improve radiologists' performance in breast cancer screening, 2020.

Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging, 2019.

Seong Ho Park and Kyunghwa Han. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction, 2003.

Eduardo H P Pooch, Pedro L. Ballester, and Rodrigo C. Barros. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification, 2020.

Mirjam Pot, Nathalie Kieusseyan, and Barbara Prainsack. Not all biases are bad: equitable and inequitable biases in machine learning and radiology, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2017.

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, 2021.

Schrasing Tong and Lalana Kagal. Investigating bias in image classification using model explanations, 2020.

Gael Varoquaux and Veronika Cheplygina. How i failed machine learning in medical imaging shortcomings and recommendations, 2021.

K. Winkler, F. Toberer C. Fink, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, and et a W. Stolz. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition, 2019.

Adam Yala, Peter G Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay et a. Toward robust mammography-based models for breast cancer risk, 2021.

Alice C. Yu and John Eng. One algorithm may not fit all: How selection bias affects machine learning performance, 2020.

R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, , and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, 2018.

Qingyu Zhao, Ehsan Adeli, and Kilian M. Pohl. Training confounder-free deep learning models for medical applications, 2020.

Appendix A. DICOM metadata definitions

DICOM metadata is used to analyze the distributions of the mammographies in the dataset to explore posible biases. The definitions of theese different tags are the following:

- **Manufacturer**: manufacturer of the equipment that produced the images.
- **Manufacturer Model Name**: model of the equipment that is used.
- **Institution Name**: institution or organization where the study was performed.
- **Detector ID**: the ID or serial number of the detector used to acquire the image.
- **Image Type**: identifies important image identification characteristics regarding the pixel data, patient examination, modality and the implementation specific identifiers.

- **Window Center/Window Width**: window center and width specify a linear conversion (unless otherwise specified by the value of VOI LUT Function) from the output of the (conceptual) Modality LUT values to the input (conceptual) Presentation LUT.

- **VOI LUT Function**: the VOI LUT Function specifies a potentially non-linear conversion to apply to the values of the image based on the window center and window width.

- **Presentation LUT Shape**: when present, specifies an identity transformation for the Presentation LUT such that the output of all grayscale transformations, if any, are defined to be in P-Values.

- **Presentation Intent Type**: identifies the intent for the purposes of display or other presentation of all images within the series.

Appendix B. Images

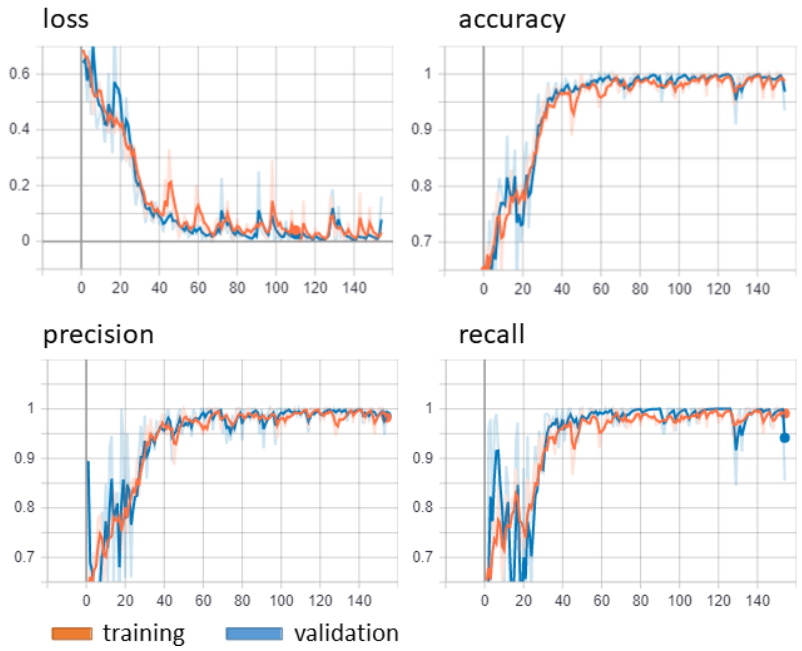

Figure 3: Loss function, accuracy, precision and recall curves during the training of the classification experiment network.

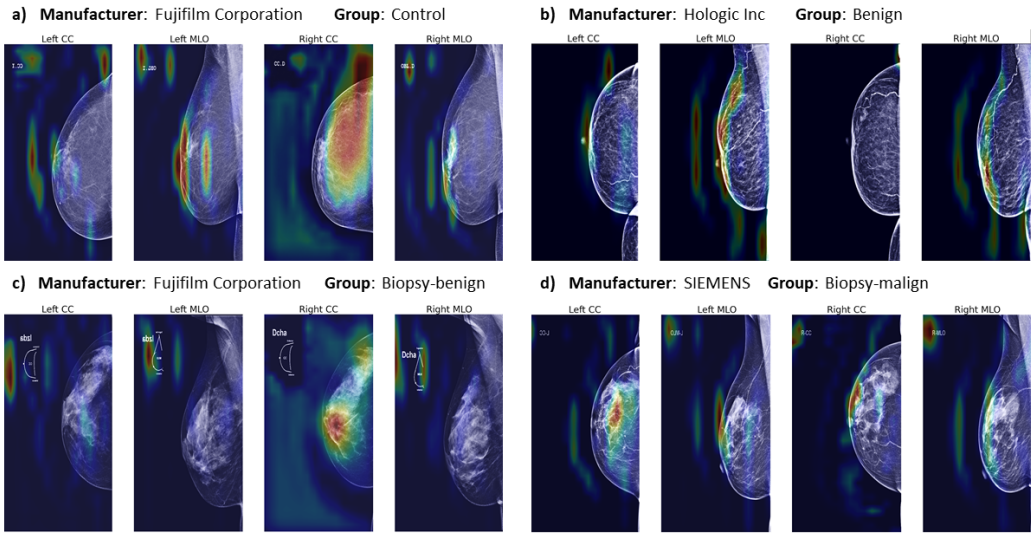

Figure 4: Grad-cam computed for the test set with the trained screeening classification experiment. Visualization of left and right breasts and different views (CC and MLO) for the different clinical categories and manufacturers is shown

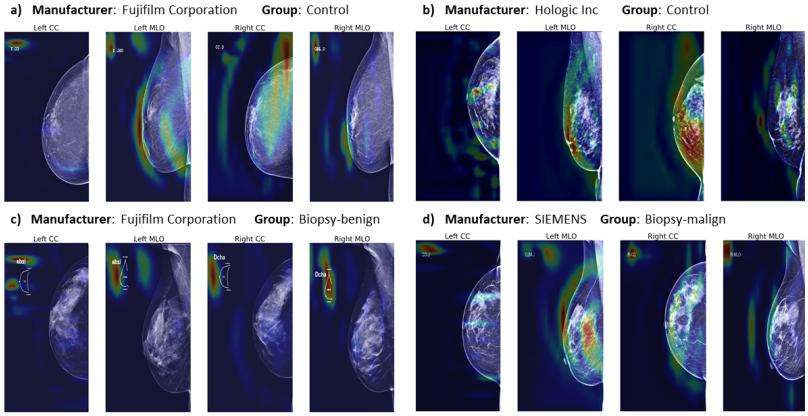

Figure 5: Grad-cam computed on the test for the device classification experiment. Left and Right breasts and different views (CC and MLO) for the different subgroups and different manufacturers are displayed. The algorithm focusses on characteristics of the image types specific of each acquisition device. Specially, the background labels indicating the breast and view (CC/MLO) information is one of the principal characteristics to classify images according to the device.

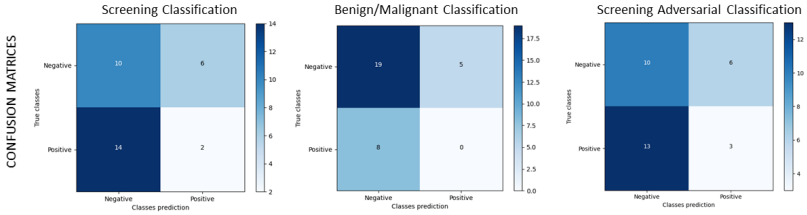

Figure 6: Confusion matrices and ROC curves for the different classification experiments on the test with the prospective set. Negative screening (Control and Benign) and positive screening (Biopsy-benign and Biopsy-malignant) studies extracted from the screening PACS were tested. In general, negative studies that were missclassified came from different devices than the ones contemplated during the training or from Philips studies with no background labels. Thus, the data bias present in the classification model is confirmed

.

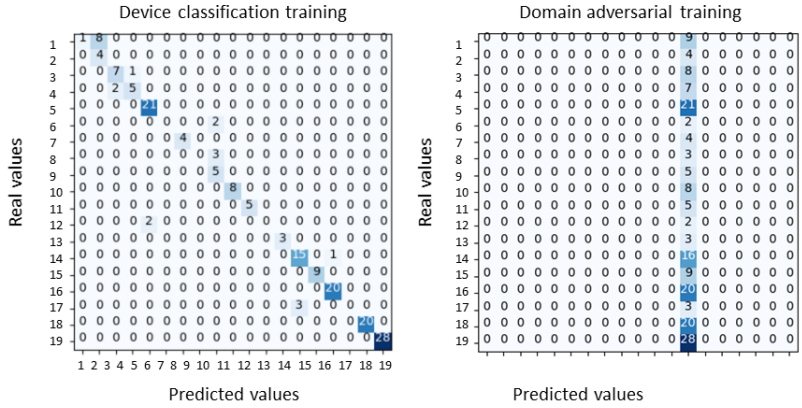

Figure 7: Confusion matrices of the device classifier (left) and the domain classifier during the domain adversarial training (right).

Appendix C. Tables

Table 5: Distribution of mammograms in the prospective test set. Number of studies of the 4 clinical categories and 4 manufacturers

	Fujifilm	**Hologic**	**Philips**	**Siemens**	**Total**
Control/Benign	3	4	5	4	16
Biopsy-benign	1	0	3	4	8
Biopsy-malignant	4	0	2	2	8
Total	8	4	10	10	32

Appendix D. Equations

Data augmentation equation applied on the fly during the classification experiments training.

$$image \equiv min\{brightness \times image + contrast, 1\} \tag{1}$$

being

$$brightness \equiv 1 + e \tag{2}$$

and "e" is randomly taken from a normal distribution with a standard deviation of 0.1. The contrast is randomly taken from a gamma distribution with alpha=1 and beta=10.