IMPACT OF IMAGE AND FEATURE HARMONIZATION ON THE COMPUTERIZED ANALYSIS OF MAMMOGRAMS

JUAN SEBASTIAN GUERRERO PEÑA
RAFAEL SANTIAGO SUÁREZ GIL

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
INGENIERÍA ELECTRÓNICA
BUCARAMANGA
2024

IMPACT OF IMAGE AND FEATURE HARMONIZATION ON THE COMPUTERIZED ANALYSIS OF MAMMOGRAMS

JUAN SEBASTIAN GUERRERO PEÑA
RAFAEL SANTIAGO SUÁREZ GIL

A dissertation submitted in partial fulfillment of the requirements for the degree of electronic engineer

Advisors:
SAID DAVID PERTUZ ARROYO
ELECTRONIC ENGINEER. PhD.
ANGIE NICOLE HERNÁNDEZ DURÁN
PHYSICIST. MSc.

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA Y DE TELECOMUNICACIONES
INGENIERÍA ELECTRÓNICA
BUCARAMANGA
2024

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my parents, Edilma and Evaristo, For supporting and encouraging me throughout my academic and professional journey. Their belief in me has been a constant source of motivation. I am deeply thankful to Jennifer, for being my steadfast companion and confidante throughout this academic endeavor. Her unwavering support and understanding have made this journey a meaningful and memorable experience. I extend my sincere appreciation to our advisors, Said and Nicole, for their invaluable guidance, support, and mentorship. Their expertise and dedication have been instrumental in shaping this work. I also want to acknowledge my thesis colleague, Rafael, for being an excellent companion in work and a great friend. His collaboration and support have been truly invaluable. Lastly, I would like to thank all those who have supported and encouraged me along the way. Thank you.

-Juan Sebastian Guerrero Peña

I extend my deepest gratitude to my loving parents, Rafael Orlando and Janneth, who have been unwavering sources of support and wise guidance throughout my life's journey. Immense gratitude goes to our esteemed advisors Said and Nicole, whose willingness to generously share their vast knowledge and experience made this work possible. Special thanks to my dedicated colleague Juan Sebastian, who proved to be not only an excellent companion but also a great mentor and a true friend. Lastly, I express my gratitude to the *CPS research group* biomedical imaging team and all those who offered their encouragement and heartfelt wishes.

-Rafael Santiago Suárez Gil

Dedicated to our families and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ANNEXES

RESUMEN

TÍTULO: IMPACTO DE LA ARMONIZACIÓN DE IMÁGENES Y CARACTERÍSTICAS EN EL ANÁLISIS COMPUTARIZADO DE MAMOGRAMAS. *

AUTORES: JUAN SEBASTIAN GUERRERO PEÑA, RAFAEL SANTIAGO SUÁREZ GIL **

PALABRAS CLAVE: ARMONIZACIÓN, ANÁLISIS PARÉNQUIMATOSO, CÁNCER DE SENO, ANÁLISIS COMPUTARIZADO DE MAMOGRAFÍA.

DESCRIPCIÓN:

Uno de los principales desafíos que enfrentan los algoritmos de análisis de imágenes médicas es la reducida capacidad de generalización debido a las diferencias entre los conjuntos de datos utilizados para el desarrollo y los utilizados para pruebas externas. En el análisis automatizado de mamografías para la investigación del cáncer de seno, este problema se ha abordado mediante la incorporación de *técnicas de armonización* de preprocesamiento de imágenes o posprocesamiento de características. Esta investigación tiene como objetivo evaluar el impacto de las técnicas de armonización en el rendimiento de los algoritmos de aprendizaje automático para la evaluación del riesgo de cáncer de seno. Realizamos un estudio retrospectivo de casos y controles sobre 147 mamografías adquiridas con sistemas mamográficos de dos proveedores diferentes y consideramos el impacto de las técnicas de armonización en tres entornos experimentales diferentes: 1) Intra-proveedor, cuando tanto las imágenes de entrenamiento como las de prueba se capturan con sistemas del mismo proveedor 2) Inter-proveedor, cuando las imágenes de entrenamiento y prueba se capturan con sistemas de diferentes proveedores y 3) Proveedor mixto, cuando el entrenamiento y las pruebas incluyen imágenes de ambos proveedores. Nuestros resultados muestran que el preprocesamiento de imágenes puede tener un impacto perjudicial en el rendimiento en entornos intra-proveedor; el uso conjunto del preprocesamiento de imágenes y el posprocesamiento de características es beneficioso en entornos inter-proveedor; y el entorno de proveedor mixto no parece verse afectado significativamente por la armonización.

---

* Tesis de Pregrado

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Directores: Said David Pertuz Arroyo. PhD. & Angie Nicole Hernández Durán. Mg

ABSTRACT

TITLE: IMPACT OF IMAGE AND FEATURE HARMONIZATION ON THE COMPUTERIZED ANALYSIS OF MAMMOGRAMS [*]

AUTHORS: JUAN SEBASTIAN GUERRERO PEÑA, RAFAEL SANTIAGO SUÁREZ GIL [**]

KEYWORDS: HARMONIZATION, PARENCHYMAL ANALYSIS, BREAST CANCER, COMPUTERIZED MAMMOGRAPHY ANALYSIS.

DESCRIPTION:

One of the main challenges faced by medical image analysis algorithms is the reduced generalization capability due to differences between datasets used for development, and those used for external testing. In the automated analysis of mammograms for breast cancer research, this problem has been tackled by the incorporation of image pre-processing or feature post-processing *harmonization techniques*. This research aims to assess the impact of harmonization techniques on the performance of machine learning algorithms for breast cancer risk assessment. We conduct a retrospective case-control study on 147 mammograms acquired with mammographic systems from two different vendors and consider the impact of harmonization techniques in three different experimental settings: 1) Intra-vendor, when both training and testing images are captured with systems from the same vendor, 2) Inter-vendor, when training and testing images are captured with systems from different vendors, and 3) Mixed-vendor when training and testing include images from both vendors. Our results show that image pre-processing can have a detrimental impact on performance in intra-vendor settings; the joint use of image pre-processing and feature pos-processing is beneficial in inter-vendor settings; and the mixed-vendor setting does not appear to be significantly affected by the harmonization.

---

[*] Undergraduate Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Advisors: Said David Pertuz Arroyo. PhD. & Angie Nicole Hernández Durán. MSc

INTRODUCTION

Breast cancer is the most common cancer among the adult population and ranks among the top two causes of cancer-related deaths in women in most countries [1]. Early breast cancer detection has proven crucial to reduce mortality [2]. Digital mammography is the main medical imaging technique for early breast cancer detection, known for its high accuracy, cost-effectiveness, and low radiation exposure [3]. Recent studies on computerized mammography analysis have shown that artificial intelligence-based (AI-based) models can match radiologists' performance in breast cancer detection [4] and serve as a reliable second opinion in computer aided diagnosis [5]. AI-based models often rely on the extraction of features, frequently referred to as parenchymal features, whether hand crafted (e.g. in radiomic analysis [6]), or abstract (e.g. in convolutional neural networks [4]).

Despite their potential, these AI-based models often face challenges in external vali-

---

[1] World Health ORGANIZATION. *WHO launches new roadmap on breast cancer*. URL: https://www.who.int/news/item/03-02-2023-who-launches-new-roadmap-on-breast-cancer.

[2] Ismail JATOI and Anthony B. MILLER. "Why is breast-cancer mortality declining?" In: *The Lancet Oncology* 4 (4 2003), pp. 251–254. DOI: 10.1016/S1470-2045(03)01037-4.

[3] Etta D. PISANO et al. "Diagnostic Accuracy of Digital versus Film Mammography: Exploratory Analysis of Selected Population Subgroups in DMIST". in: *Radiology* 246 (2 Feb. 2008), pp. 376–383. DOI: 10.1148/radiol.2461070200.

[4] Li SHEN et al. "Deep Learning to Improve Breast Cancer Detection on Screening Mammography". In: *Scientific Reports* 121 (2019). DOI: 10.1038/s41598-019-48995-4.

[5] I. ANDREADIS et al. "Investigating the performance of a CADx scheme for mammography in specific BIRADS categories". In: *IEEE International Conference on Imaging Systems and Techniques* (Nov. 2014), pp. 335–339. DOI: 10.1109/IST.2014.6958500.

[6] Said PERTUZ et al. "Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: A pilot study in a Finnish sample". In: *European Journal of Radiology* 121 (2019). DOI: 10.1016/j.ejrad.2019.108710.

dation, where image analysis is conducted on unseen data [7]. By design, the external validation dataset often consist of images acquired under conditions that differ from those of the training and testing datasets; in mammography, this often implies images acquired with different mammographic systems. The technologies and image enhancement algorithms used by different vendors result in mammograms that show system-inherent, non-biological characteristics [8]. In turn, these characteristics affect algorithm performance and reproducibility [9][10].

Mitigating the negative effects on model performance due to different imaging acquisition conditions is one of the current challenges of mammography analysis. This challenge directly relates to the concept of model robustness, furthermore, using features that are robust among manufacturers can lead to more robust models [11][12]. In this work, we term the techniques aimed at improving model robustness as *harmonization* techniques, which we categorize in two groups: image pre-processing and feature post-processing.

---

[7] X. WANG et al. "Inconsistent Performance of Deep Learning Models on Mammogram Classification". In: *J Am Coll Radiol.* 17.6 (2020), pp. 796–803.

[8] Yan WANG et al. "A phantom study for assessing the effect of different digital detectors on mammographic texture features". In: *Breast Imaging*. Springer, 2012, pp. 604–610.

[9] Kayla R. MENDEL et al. "Quantitative texture analysis: robustness of radiomics across two digital mammography manufacturers' systems". In: *Journal of Medical Imaging* 5 (1 2018), p. 011002. DOI: 10.1117/1.JMI.5.1.011002.

[10] Said PERTUZ et al. "Do Mammographic Systems Affect the Performance of Computerized Parenchymal Analysis?" In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2019, pp. 4863–4866. DOI: 10.1109/EMBC.2019.8856948.

[11] Raymond J. ACCIAVATTI et al. "Incorporating Robustness to Imaging Physics into Radiomic Feature Selection for Breast Cancer Risk Estimation". In: *Cancers* 13 (21 Nov. 2021), p. 5497. DOI: 10.3390/cancers13215497.

[12] Kayla ROBINSON et al. "Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFDM". in: *Medical physics* 46 (5 May 2019), pp. 2145–2156. DOI: 10.1002/MP.13455.

Image pre-processing techniques imply the modification of the image pixel values prior to any analysis; this approach has been explored by Wang et al. [13], who performed a phantom study to assess the effect of mammography image normalization using z-score. They found that this approach greatly alleviated system-inherent differences in grey-level histogram features. Feature post-processing can be explored in the form of feature standardization through strategies like the aforementioned z-score, and statistical model-based approaches, such as the ComBat method [14], which has shown success in alleviating the batch effect in various scenarios for different imaging modalities [15][16].

Despite the aforementioned efforts to compensate for changes in acquisition conditions, the literature does not offer consensus regarding which harmonization techniques are the most effective to compensate for changes in mammography acquisition devices. There is a lack of comprehensive examination of the combined impact of image pre-processing and feature post-processing, therefore, there is no experimental evidence on how these strategies interact and their overall impact on computerized mammographic analysis tasks. The aim of this work is to determine how different harmonization approaches impact the performance of computerized mammographic image analysis in breast cancer risk classification. We are particularly interested in the effects of harmonization in a multi-vendor scenario. For this, we establish various har-

---

[13] Yan WANG et al. "Texture feature standardization in digital mammography for improving generalizability across devices". In: *Medical Imaging 2013: Computer-Aided Diagnosis*. SPIE, 2013, p. 867026.

[14] W Evan JOHNSON et al. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[15] J. P. FORTIN et al. "Harmonization of multi-site diffusion tensor imaging data". In: *Neuroimage* 161 (2017), pp. 149–170.

[16] Joaquim RADUA et al. "Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA". in: *Neuroimage* 218 (2020), p. 116956.

monization strategies by combining different approaches, and leverage our two-vendor dataset to train and test different models using several dataset partitions.

1. MATERIALS AND METHODS

1.1. IMAGING DATA

For this retrospective case-control study we used images from the Emory breast imaging dataset (EMBED) [17]. The freely available EMBED dataset version contains a total of 480,323 images in DICOM format linked with clinical data. In accordance with our goal, we selected data to comply with three requirements:

First, imaging data should allow performing a specific image analysis task. We opted for *breast cancer risk assessment*, where the computerized model solves a binary classification problem based on parenchymal features. For risk assessment, the model should classify each image as *high risk* for cases, or *low risk* for controls (see section 1.2). For cases, we selected images categorized as 5 or 6 in the BI-RADS malignancy assessment, indicating a likelihood of malignancy greater than 95% or confirmed malignancy. Control images were selected from BI-RADS category 1, i.e. negative for malignancy.

Second, data should include images captured with different mammography manufacturers, for the assessment of vendor-inherent effects. We selected images from two manufacturers, Hologic (vendor 1) and General Electric (vendor 2), and ensured that the imaging data included cases and controls from each manufacturer.

Third, data selection should reduce the effect of confounding variables on the analysis, such as age, breast density and ethnicity. Age is one of the main risk factors associated to breast cancer and breast density and ethnicity may act as further model confounders

---

[17] Jiwoong J. JEONG et al. "The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images". In: *Radiology: Artificial Intelligence* 5 (1 Jan. 2023). DOI: 10.1148/ryai.220047.

[18]. Therefore, we *matched* cases and controls by age, breast density category and ethnicity.

Figure 1 illustrates the complete data selection process. Image selection according to BI-RADS categories and the mammographic system was performed based on the clinical data of each examn. Images that failed matching were excluded. Due to missing views, we only selected images of the craniocaudal (CC) view for each women. We selected one single view in some exams with repeated views. Finally, all images were subject to visual inspection to discard artifacts or corrupt data. At the end of the selection process, 147 images were suitable for inclusion in our study. Table 1 summarizes the distribution of the selected images.

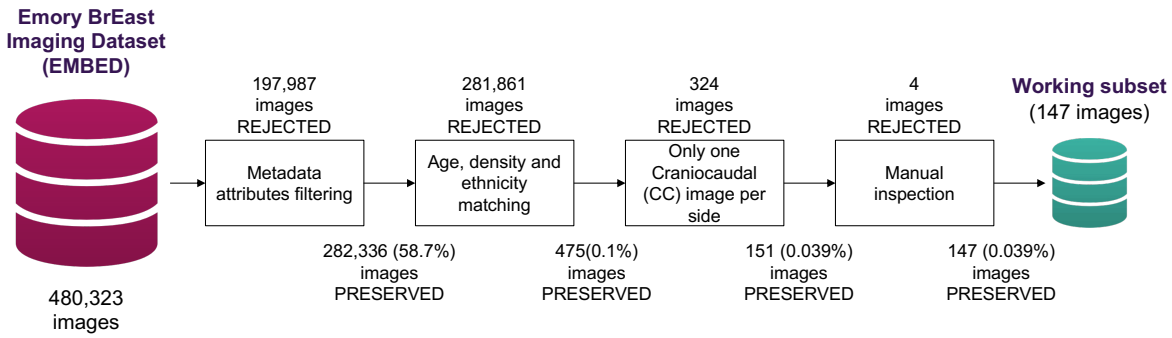

Figure 1. Selection of imaging data

Table 1. Imaging data by vendor and assessment groups

Category	Vendor 1	Vendor 2	Total
Cases	20	22	42
Controls	49	56	105
Total	69	78	147

[18] Gabriel N. HORTOBAGYI et al. "The Global Breast Cancer Burden: Variations in Epidemiology and Survival". In: *Clinical Breast Cancer* 6 (5 Dec. 2005), pp. 391–401. DOI: 10.3816/CBC.2005.N.043.

1.2. BREAST CANCER RISK ASSESSMENT

We used parenchymal analysis for the task of breast cancer risk assessment. In the literature, automatic parenchymal analysis uses high throughput computerized imaging features in order to build risk assessment models for mammographic images [6]. Following that, we used the OpenBreast [19] framework to extract a total of 33 parenchymal features from each image. Based on Pertuz et al. [10], features were extracted using the full breast region. These features encompass five different categories: statistical, run-length, co-occurrence, gradient-based and spatial-frequency features (Fig. 2). Breast cancer risk assesment was performed via logistic regression. For feature selection and regularization, we considered three techniques: Lasso regularization [20], ElasticNet regularization [21], and stepwise forward feature selection.

1.3. EXPERIMENTAL SETTINGS AND HARMONIZATION SCENARIOS

To assess features robustness, each feature distribution is compared between the two vendors using the two sample Kolmogorov-Smirnov (K-S) test at 0.05 significance [13]. The null hypothesis (H0) states that the two samples come from the same distribution. In other words, if p-value$<$0.05, there is evidence to reject the null hypothesis, i.e. the two samples are significantly different. Consequently, features with a p-value greater than 0.05 are considered robust to changes in mammography system. To assess classification performance, we used five-fold cross validation to estimate the area under

---

[19] Said PERTUZ et al. “Open Framework for Mammography-based Breast Cancer Risk Assessment”. In: *IEEE EMBS International Conference on Biomedical & Health Informatics*. 2019, pp. 1–4. DOI: 10.1109/BHI.2019.8834599.

[20] Robert TIBSHIRANI. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B* 58 (1 1996), pp. 267–288.

[21] Hui ZOU and Trevor HASTIE. “Regularization and variable selection via the elastic net”. In: *J. R. Statist. Soc. B* 67 (2 2005), pp. 301–320.

Figure 2. Radiomic feature extraction

the receiver ROC curve (AUC) with 95% confidence intervals. Because we wanted to assess the effect of harmonization strategies on model robustness to changes induced by different mammographic vendors, we considered three different experimental settings:

1. *Mixed-vendor cross validation*. Training and testing sets are randomly selected from the full dataset, regardless of mammographic vendor.

2. *Intra-vendor cross validation*. Training and testing sets only include images of the same vendor. As a result, in each validation fold we obtain two models, one for each mammographic vendor.

3. *Inter-vendor cross validation*. Training set includes images from one vendor and testing set includes images from the other vendor. We obtain two models, one for each mammographic vendor.

To simplify the terminology, we will refer to image pre-processing strategies as *im-*

age harmonization and to parenchymal texture feature post-processing strategies as *feature harmonization*. Image harmonization is applied by z-score normalization (see ANNEX A) on the graylevel values of the segmented breast region, whereas feature harmonization is applied by means of the ComBat method [14] (see ANNEX B). Due to its impact on parenchymal features [18], the breast density category is defined as a co-variate for ComBat harmonization.

As shown in Figure. 3, the inclusion of image or feature harmonization allows us to consider four different harmonization scenarios before training the prediction models. Each of the four harmonization scenarios described below are tested on the three experimental settings described previously:

1. *No harmonization*, when neither image harmonization or feature harmonization are used

2. *Feature harmonization*, when feature harmonization is applied to features extracted from the original image.

3. *Image harmonization*, when image harmonization is applied before feature extraction. The features extracted on the harmonized image are used directly for the construction of the prediction model..

4. *Image and feature harmonization*, when image harmonization is applied before feature extraction. Then, feature harmonization is applied to the extracted features.

In order to determine if two resulting AUCs (e.g. no harmonization and image harmonization resulting AUCs) are significantly different we used the *DeLong's test* [22].

---

[22] Elizabeth R. DELONG, David M. DELONG, and Daniel L. CLARKE-PEARSON. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach". In: *Biometrics* 44 (3 Sept. 1988), p. 837. DOI: 10.2307/2531595.

The null hypothesis states that both AUCs are not significantly different. If p-value $<$ 0.05, there is evidence to reject the null hypothesis, i.e. suggest that the two AUCs are significantly different.

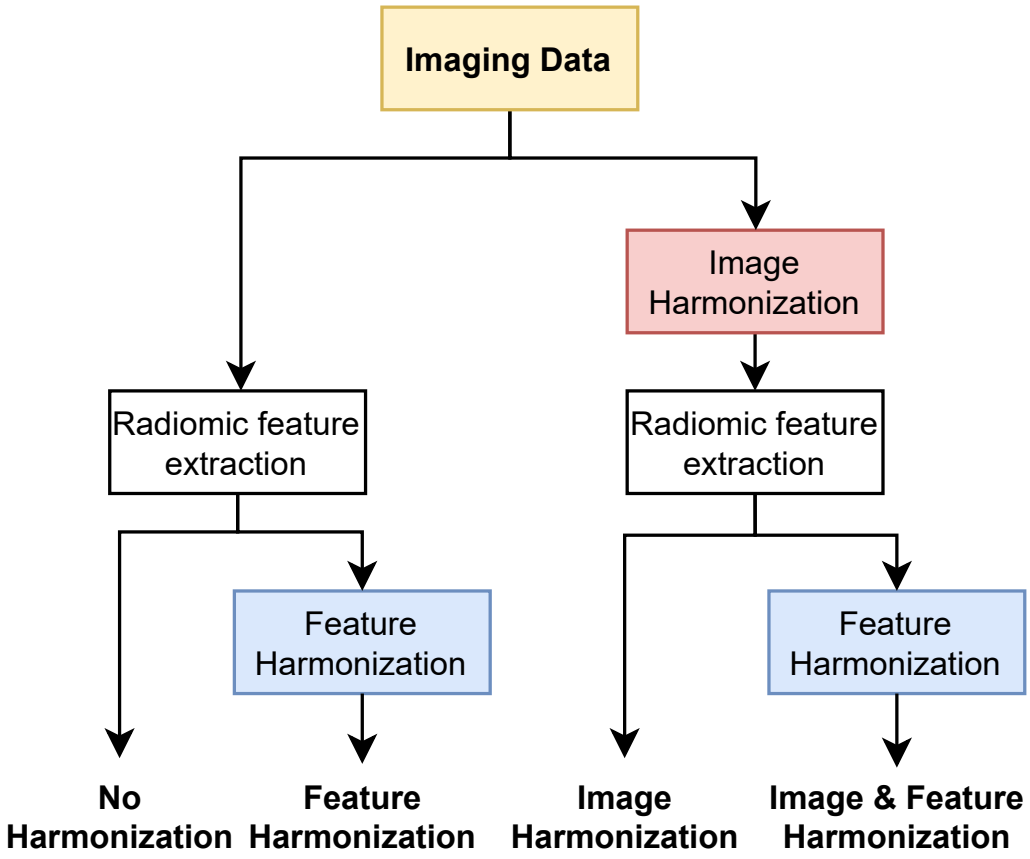

Figure 3. Harmonization scenarios

2. RESULTS AND DISCUSSION

2.1. EFFECTS OF HARMONIZATION ON TEXTURE FEATURES

Our motivation for employing harmonization techniques was to mitigate differences in parenchymal features caused by variations in imaging conditions. Using the two sample K-S test we determined the impact of the 4 harmonization scenarios on the difference of the pairs of distributions of each texture feature. Table 2 shows the p-values of robust features ($p > 0.05$) for each harmonization scenario.

According to Table 2, only two texture features are inherently robust to variations in imaging conditions and this number increases when applying harmonization techniques. The scenario with the least improvement is image harmonization which only added two robust features, both from the statistical category. This makes sense since these two features, mean gray level value (savg) and gray level variance (svar), are directly affected by the process of standardizing gray level values. Although there are features distributions that shows no discernible impact when image harmonization is applied, as in the case of gray level entropy (sent) shown in figure 4, there are cases in which the defined significance (0.05) is not exceeded, but a considerable impact could be verified visually. This was the case, for instance, of gradient-based and spatial-frequency features. Fig. 5 show the example of the gradient energy feature.

The scenario that enhances robustness the most is the one that combines image and feature harmonization. This was expected since, in addition to bringing the gray levels to a standardized distribution before extracting features, a process is applied to adjust the two distributions of each feature after extracting features.

Table 2. Results of comparison of features between both vendors
Results of comparison of features between both vendors by two-sample Kolmogorov–Smirnov test on each of the harmonization scenarios. Feature distributions with p-values$>$0.05 are shown and considered robust. Total is the count of features considered robust.

Feature	**K-S test p-values**			
	Harmonization			
Abbr.	**None**	**Feature**	**Image**	**Image&feature**
smin				0.457
smax				0.560
savg			1.000	1.000
sran				0.676
svar			1.000	1.000
sent		0.815		0.722
sske		0.221		0.221
skur				
sp05				0.705
sp30		0.073		0.659
sp70				0.739
sp95				0.778
sba1		0.928		0.928
sba2		0.756		0.756
cene		0.154		0.288
ccor	0.074	0.705	0.074	0.572
ccon		0.084		
chom		0.145		0.090
cent		0.688		0.688
rrln	0.241	0.606	0.241	0.494
rgln		0.659		0.659
rlre				0.059
rsre		0.152		0.100
rrpe		0.152		0.105
rhgr		0.103		0.114
rlgr				
gene				0.859
gvar				0.560
glap				0.572
fwas				0.538
fwav				0.583
fwar				0.699
fdim		0.516		0.421
Total	2	16	4	30

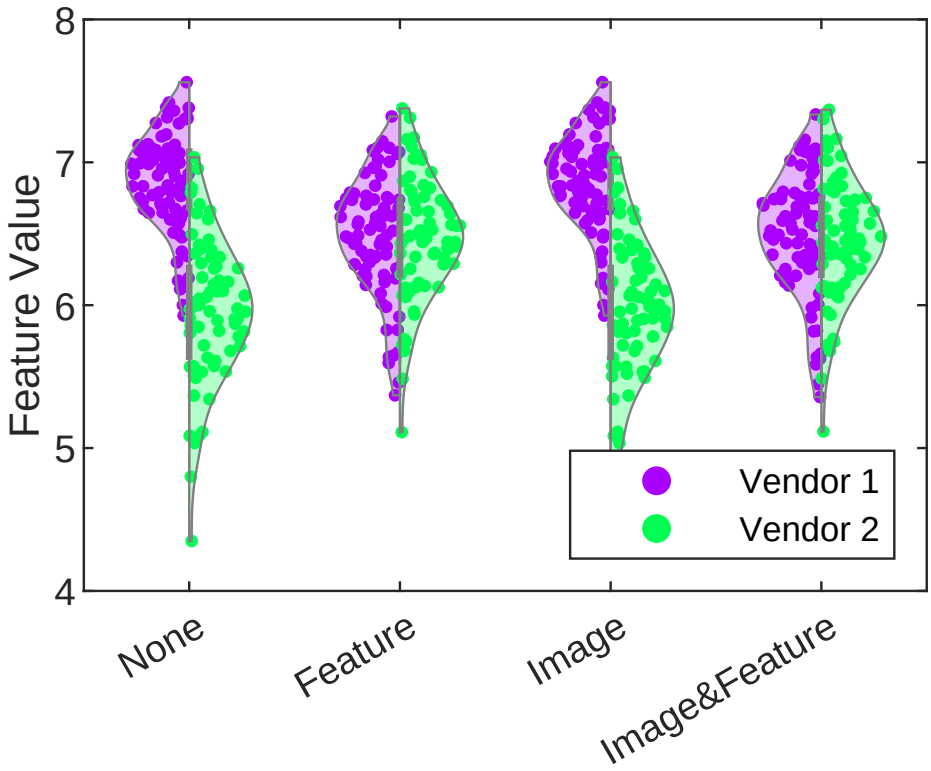

Figure 4. Effect of harmonization on the gray level entropy (sent) feature.
In this case, the feature distributions start off with a notorious difference in means; this difference seems to be greately alleviated by feature harmonization, and not too affected by image harmonization.

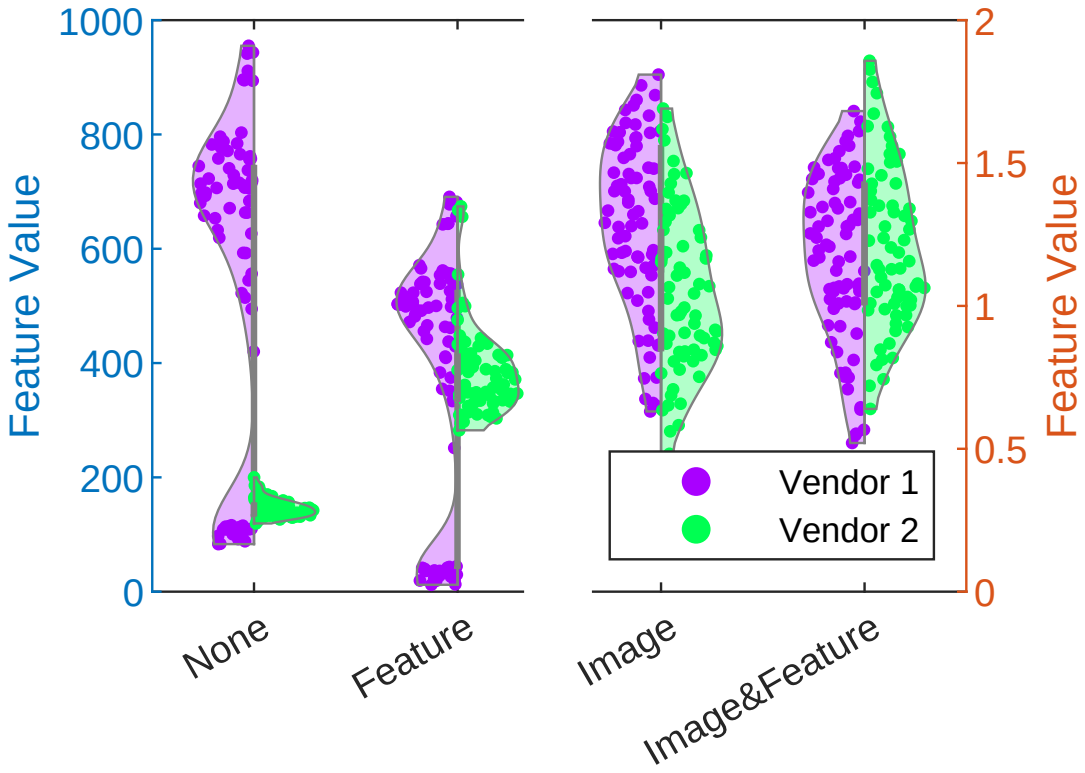

Figure 5. Effect of harmonization in the gradient energy (gene) feature.
In this case, the feature distributions start off drastically different, and the effect of feature harmonization does not seem to effectively alleviate for this. Image harmonization, however, does have a notable effect, so much so that the original scale is drastically changed.

2.2. HARMONIZATION EFFECTS ON CLASSIFICATION

The AUCs of breast cancer risk assessment for all the experimental settings and harmonization scenarios are presented in Table 3. In that table, significantly different AUC values according to DeLong test (p-value$<$0.05) compared to *no harmonization* setting are indicated by color, blue for performance improvements and red for performance deterioration. The intra-vendor setting represents an idealistic scenario in which the model is trained and tested with data acquired with the same vendor. In this case, the model does not face unseen system-inherent effects, which results in a significant positive performance without need for harmonization. Notably, the best AUC results were obtained using mixed vendors, which were also not significantly affected by harmonization. In general, the lowest performances are observed in the inter-vendor with no harmonization settings. This is not surprising, since this is the most challenging setting with training and testing images being captured with systems from different vendors.

Table 3. AUCs for different feature selection and regularization strategies.
AUCs for different feature selection and regularization strategies. Significantly different AUC values (p-value$<$0.05) compared to no harmonization setting are indicated by color, blue for performance improvements and red for performance deterioration.

	Harmonization	Inter-vendor	Intra-vendor	Mixed-vendors
ElasticNet	None	0.60 (0.50-0.71)	0.73 (0.64-0.83)	0.77 (0.68-0.86)
	Feature	0.61 (0.50-0.71)	0.72 (0.62-0.81)	0.79 (0.71-0.88)
	Image	0.72 (0.62-0.82)	0.60 (0.50-0.71)	0.74 (0.64-0.83)
	Image&Feature	0.76 (0.67-0.86)	0.61 (0.51-0.71)	0.73 (0.64-0.83)
Lasso	None	0.57 (0.46-0.67)	0.72 (0.62-0.82)	0.77 (0.68-0.86)
	Feature	0.65 (0.55-0.75)	0.71 (0.61-0.81)	0.79 (0.70-0.88)
	Image	0.74 (0.65-0.84)	0.61 (0.50-0.71)	0.76 (0.67-0.85)
	Image&Feature	0.75 (0.66-0.85)	0.63 (0.53-0.73)	0.74 (0.64-0.83)
Stepwise	None	0.64 (0.54-0.74)	0.75 (0.66-0.84)	0.67 (0.57-0.78)
	Feature	0.71 (0.61-0.81)	0.75 (0.66-0.84)	0.77 (0.68-0.86)
	Image	0.74 (0.65-0.84)	0.71 (0.61-0.81)	0.68 (0.58-0.78)
	Image&Feature	0.72 (0.63-0.82)	0.72 (0.62-0.82)	0.63 (0.53-0.74)

To facilitate discussion according to the experimental setting, the results of Table 3 using Lasso regularization are illustrated in Fig. 6. In that Figure, AUCs are represented with a marker and confidence intervals are shown as whiskers. For a result to be statistically significant, whiskers should be above the 0.5 value (dashed vertical line). As shown in that figure, the effects of the harmonization approach in the performance of automatic risk classification yield mixed results depending on the experimental setting.

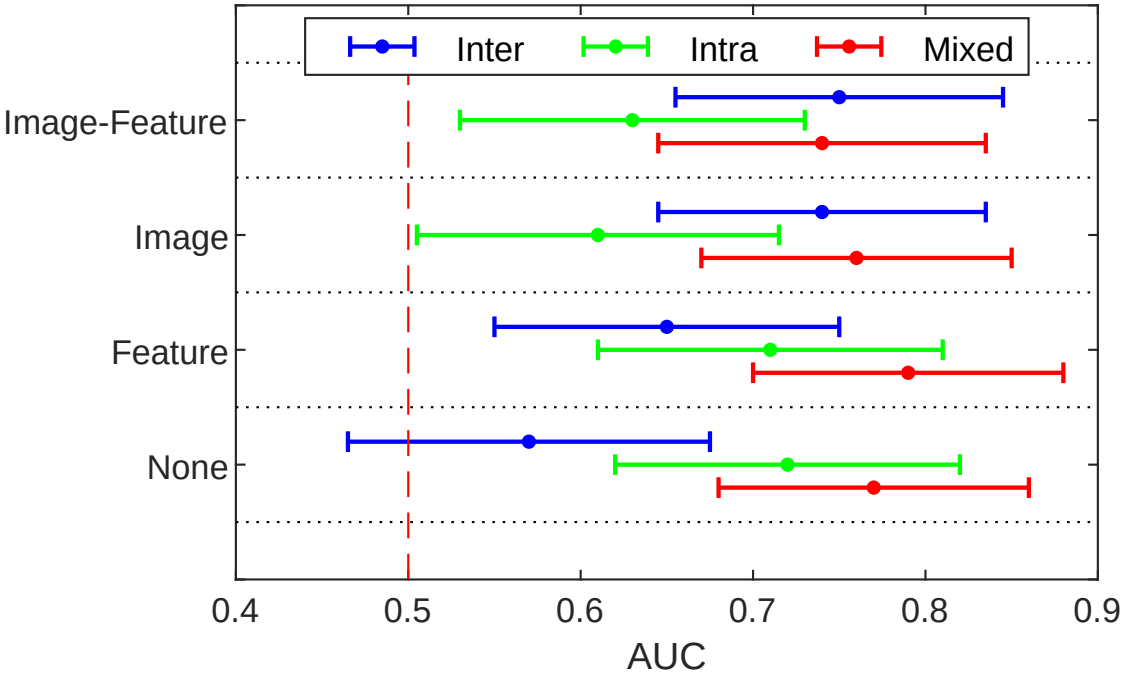

Figure 6. AUC for all experimental settings
AUC in breast cancer risk assessment for all experimental settings and harmonization scenarios using Lasso

The clearest effect of harmonization is observed in the inter-vendor setting. In such, the model progresses from a lack of statistically significant results when no harmonization is used, to achieving significant performances with harmonization. Specifically, performance increases when feature harmonization, image harmonization, and image & feature harmonization are applied, in that order. The inter-vendor scenario is the most

demanding experimental setting since the prediction model is trained using images from one vendor and tested in images from a different vendor. Our results suggest that model generalization in this setting is achieved by using both image and feature harmonization. The relevance of these results lays in the fact that, in a realistic setting, it is not guaranteed that the data used for testing or validation will reflect the same characteristics as the training data. In this scenario, the model faces data with unseen characteristics yet, due to the harmonization strategies, it is able to achieve a significant performance.

In the intra-vendor scenario, harmonization either has no effect or worsens classification results. The lowest performance is achieved when using image harmonization only. Since train and test images are captured with systems from the same vendor in this scenario, we argue that imaging data already shares the same non-biological characteristics due to acquisition. Therefore, the harmonization procedures are, in the best case, not affecting the image characteristics meaningfully and, in the worst case, reducing the model's classification performance.

In the mixed vendor scenario, there are no significant differences in performance. These results suggest that harmonization does not have a positive or negative impact on the classification performance of the model. These findings are highly relevant since, in practice, it is likely that computerized image analysis models should be designed to work in this experimental setting.

This work has two main limitations. First, the reduced data sample (N=147) results in large confidence intervals when estimating the AUCs. This adds uncertainty when comparing small differences in performance between the experimental settings and harmonization scenarios. The contrast between the resulting small sample size after our image selection process and the size of the original EMBED dataset puts into perspective the difficulties in conducting this type of studies. Second, our analysis is limited to feature-based machine learning classification algorithms. In the litera-

ture, CNNs have shown great promise in mammographic image analysis. However, mammography data or image harmonization strategies in this scope remain an unexplored problem, whereas generalizability is often tackled through domain adaptation and transfer learning techniques.

3. CONCLUSIONS

We studied the effectiveness of image pre-processing and feature post-processing strategies, namely harmonization, for improving the generalizability of computerized mammographic analysis algorithms. For this purpose, we conducted a restrospective case-control study on 147 mammograms for the task of breast cancer risk assessment from parenchymal features. Our results show that mixing image and feature harmonization techniques makes features more robust to changes in imaging conditions, however, this robustness does not necessarily translate into model performance improvement, and it is very dependent on the three different experimental settings tested. In the most challenging case, with training and testing images acquired from different sources, using harmonization strategies has a significant improvement in model performance. On the other hand, in the most practical case, in which both training and testing data mix images from different acquisition sources, using harmonization strategies has not a significant impact in the model performance.

4. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using open access human subject data from *EMBED* dataset [17] . Ethical approval was not required.

BIBLIOGRAPHY

ACCIAVATTI, Raymond J. et al. “Incorporating Robustness to Imaging Physics into Radiomic Feature Selection for Breast Cancer Risk Estimation”. In: *Cancers* 13 (21 Nov. 2021), p. 5497. DOI: 10.3390/cancers13215497 (cit. on p. 13).

ANDREADIS, I. et al. “Investigating the performance of a CADx scheme for mammography in specific BIRADS categories”. In: *IEEE International Conference on Imaging Systems and Techniques* (Nov. 2014), pp. 335–339. DOI: 10.1109/IST.2014.6958500 (cit. on p. 12).

DELONG, Elizabeth R., David M. DELONG, and Daniel L. CLARKE-PEARSON. “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach”. In: *Biometrics* 44 (3 Sept. 1988), p. 837. DOI: 10.2307/2531595 (cit. on p. 20).

FORTIN, J. P. et al. “Harmonization of multi-site diffusion tensor imaging data”. In: *Neuroimage* 161 (2017), pp. 149–170 (cit. on p. 14).

HORTOBAGYI, Gabriel N. et al. “The Global Breast Cancer Burden: Variations in Epidemiology and Survival”. In: *Clinical Breast Cancer* 6 (5 Dec. 2005), pp. 391–401. DOI: 10.3816/CBC.2005.N.043 (cit. on pp. 17, 20).

JATOI, Ismail and Anthony B. MILLER. “Why is breast-cancer mortality declining?” In: *The Lancet Oncology* 4 (4 2003), pp. 251–254. DOI: 10.1016/S1470-2045(03)01037-4 (cit. on p. 12).

JEONG, Jiwoong J. et al. “The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images”. In: *Radiology: Artificial Intelligence* 5 (1 Jan. 2023). DOI: 10.1148/ryai.220047 (cit. on pp. 16, 31).

JOHNSON, W Evan et al. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (2007), pp. 118–127 (cit. on pp. 14, 20, 36).

MENDEL, Kayla R. et al. “Quantitative texture analysis: robustness of radiomics across two digital mammography manufacturers’ systems”. In: *Journal of Medical Imaging* 5 (1 2018), p. 011002. DOI: 10.1117/1.JMI.5.1.011002 (cit. on p. 13).

ORGANIZATION, World Health. *WHO launches new roadmap on breast cancer*. URL: https://www.who.int/news/item/03-02-2023-who-launches-new-roadmap-on-breast-cancer (cit. on p. 12).

ORLHAC, Fanny et al. “A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies”. In: *Journal of Nuclear Medicine* 63 (2 Feb. 2022), pp. 172–179. DOI: 10.2967/JNUMED.121.262464 (cit. on p. 36).

PERTUZ, Said et al. “Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: A pilot study in a Finnish sample”. In: *European Journal of Radiology* 121 (2019). DOI: 10.1016/j.ejrad.2019.108710 (cit. on pp. 12, 18).

— “Do Mammographic Systems Affect the Performance of Computerized Parenchymal Analysis?” In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2019, pp. 4863–4866. DOI: 10.1109/EMBC.2019.8856948 (cit. on pp. 13, 18).

— “Open Framework for Mammography-based Breast Cancer Risk Assessment”. In: *IEEE EMBS International Conference on Biomedical & Health Informatics*. 2019, pp. 1–4. DOI: 10.1109/BHI.2019.8834599 (cit. on p. 18).

PISANO, Etta D. et al. “Diagnostic Accuracy of Digital versus Film Mammography: Exploratory Analysis of Selected Population Subgroups in DMIST”. In: *Radiology* 246 (2 Feb. 2008), pp. 376–383. DOI: 10.1148/radiol.2461070200 (cit. on p. 12).

RADUA, Joaquim et al. “Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA”. In: *Neuroimage* 218 (2020), p. 116956 (cit. on p. 14).

ROBINSON, Kayla et al. “Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFDM”. In: *Medical physics* 46 (5 May 2019), pp. 2145–2156. DOI: 10.1002/MP.13455 (cit. on p. 13).

SHEN, Li et al. “Deep Learning to Improve Breast Cancer Detection on Screening Mammography”. In: *Scientific Reports* 121 (2019). DOI: 10.1038/s41598-019-48995-4 (cit. on p. 12).

TIBSHIRANI, Robert. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B* 58 (1 1996), pp. 267–288 (cit. on p. 18).

WANG, X. et al. “Inconsistent Performance of Deep Learning Models on Mammogram Classification”. In: *J Am Coll Radiol.* 17.6 (2020), pp. 796–803 (cit. on p. 13).

WANG, Yan et al. “A phantom study for assessing the effect of different digital detectors on mammographic texture features”. In: *Breast Imaging*. Springer, 2012, pp. 604–610 (cit. on p. 13).

— “Texture feature standardization in digital mammography for improving generalizability across devices”. In: *Medical Imaging 2013: Computer-Aided Diagnosis*. SPIE, 2013, p. 867026 (cit. on pp. 14, 18, 35).

ZOU, Hui and Trevor HASTIE. “Regularization and variable selection via the elastic net”. In: *J. R. Statist. Soc. B* 67 (2 2005), pp. 301–320 (cit. on p. 18).

ANNEXES

ANNEX A. Z-SCORE NORMALIZATION

Z-score is a normalization method that allows to take a distribution (ideally normal) and apply a linear transformation, the goal is to fit the distribution to a standard normal distribution [13]. The equation is:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

Where X is the original data, μ is the arithmetic mean and σ is the standard deviation.

ANNEX B. COMBAT HARMONIZATION

ComBat is a feature harmonization method [14]. ComBat assumes that the difference between the "sites" is produced by an additive and multiplicative factor. The equation is as follows:

$$y_{ij} = \alpha + \gamma_i + \delta_i \epsilon_{ij} \tag{2}$$

Where i is the system, j is the image of interest, y is the feature, α is the arithmetic mean value of the feature, γ is the additive effect,δ_i is the multiplicative effect and ϵ_{ij} is an error term. By estimating γ and δ_i it is possible to correct the effect with:

$$y_{ij}{}^{ComBat} = \frac{y_{ij} - \hat{\alpha} - \hat{\gamma_i}}{\hat{\delta_i}} + \hat{\alpha} \tag{3}$$

where α, γ_i and δ_i are estimated as $\hat{\alpha}$, $\hat{\gamma_i}$ and $\hat{\delta_i}$. ref: [23]

[23] Fanny ORLHAC et al. "A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies". In: *Journal of Nuclear Medicine* 63 (2 Feb. 2022), pp. 172–179. DOI: 10.2967/JNUMED.121.262464.