Contrastive learning: an efficient Domain Adaptation strategy for 2D mammography image classification

Gonzalo Iñaki Quintana, Vincent Jugnon, Laurence Vancamberg, Agnès Desolneux, Mathilde Mougeot

CONTRASTIVE LEARNING: AN EFFICIENT DOMAIN ADAPTATION STRATEGY FOR 2D MAMMOGRAPHY IMAGE CLASSIFICATION

Gonzalo Iñaki Quintana[⋆†] *Vincent Jugnon*[⋆] *Laurence Vancamberg*[⋆] *Agnès Desolneux*[†] *Mathilde Mougeot*[†]

[⋆] GE HealthCare, 283 Rue de la Minière, 78530 Buc, France
[†] ENS Paris-Saclay, Centre Borelli, F-91190 Gif-sur-Yvette, France

ABSTRACT

Effective computer aided breast-cancer diagnosis models using 2D mammography images must maintain consistent performance across varying image acquisition systems and post-processing techniques. Nevertheless, Deep Learning (DL) models have shown diminished performance with variations in image style and contrast [1, 2]. We propose two models trained for classifying respectively 2D mammography patches and complete images, using heterogeneous datasets distinguished by different image post-processing methods. We propose a Domain Adaptation (DA) methodology using Supervised Contrastive Learning (SCL) to achieve domain-invariant representations and improved class-separability. This approach is compared to a standard training using the Cross Entropy (CE) loss. The domain invariant models outperform those trained with CE in binary classification of full mammograms (cancer vs. no cancer), increasing the AUC from 0.745 to 0.816 in an independent test set. For patch classification, we show that the Domain Adaptation effectiveness varies with weight initialization and dataset size.

Index Terms— Breast cancer, Deep Learning, Computer Aided Detection or Diagnosis (CAD), Domain Adaptation, Contrastive Learning

1. INTRODUCTION

X-ray imaging techniques, like Full Field Digital Mammography (FFDM), also known as 2D mammography, are widely used for breast cancer screening, but suffer from low specificity and high patient recall rates. Deep Learning (DL)-based Computer Aided Detection or Diagnosis (CAD) systems emerge today as one of the most promising techniques for improving the overall effectiveness of mammography as a screening tool for cancer detection. The development of CAD systems that match or surpass radiologists' performance is a challenging task, due to the high variability of mammography images. Moreover, the existence of multiple acquisition systems and post-processing algorithms, including variations introduced by different vendors and their product versions, adds another layer of variability that directly affects image style and contrast. It has been shown that performance strongly degrades when a CAD is tested on images with different style or contrast to those used for training [3].

1.1. Domain Adaptation

The work presented here is focused on the development of a FFDM classifier that is invariant to image post-processing differences. This is modeled as a Domain Adaptation (DA) problem, in which two or more domains are adapted during training. This formulation is slightly different from the classical Domain Adaptation setting, in which there is a source domain from which we intend to do Transfer Learning to a target domain. Let $X \in \mathcal{X}$ denote an input or covariate (in our case, a mammography image), and $Y \in \mathcal{Y}$ its associated label, with a joint probability density $P : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. The two domains considered in this work are then formalized as $\mathcal{D}_1 = \{\mathcal{X} \times \mathcal{Y}, P_1\}$ and $\mathcal{D}_2 = \{\mathcal{X} \times \mathcal{Y}, P_2\}$. If $P_1(X,Y) \neq P_2(X,Y)$, there is a *domain shift*. The image post-processing variability can be modeled as a *hidden covariate shift*, as defined in [4], also known as *covariate observation shift* [5]. It is a particular case of the *concept shift*, which occurs when $P_1(X|Y) \neq P_2(X|Y)$, for which it is assumed that there exists a linear transformation of the covariates $\phi : \mathcal{X} \to \mathcal{X}$ under which the shift is non-existent, i.e., $P_1(\phi(X),Y) = P_2(\phi(X),Y)$. For more details on the types of domain shift, refer to [4, 5, 6, 7].

In the Domain Adaptation literature, feature-based methods are used to tackle the hidden covariate shift by learning a feature representation in which the domains are indistinguishable [4]. That is, to learn a function $\phi : \mathcal{X} \to \mathcal{X}$ such that $P_1(\phi(X),Y) = P_2(\phi(X),Y)$. The most widely used feature-based DA method is Discriminative Adversarial Neural Network (DANN) [8], in which an encoder is trained to fool a domain discriminator, while providing features that are useful for the classification task. The three networks (encoder, classifier, and discriminator) are optimized by jointly minimizing the classifier loss and maximizing the discriminator loss. DANN has already been applied to the medical imaging domain [9, 10, 11], and has shown promising results. However, this approach needs an additional network (the domain discriminator), which increases training time and resource consumption, mostly in terms of GPU RAM. This can be especially problematic for some medical imaging applications, like classification and detection in mammography images, in which reducing the image resolution can highly impact performance [12]. In addition, the gradients of the two losses usually have different directions, which makes DANNs hard to train [13]. Another widely used approach consists in minimizing a dissimilarity measure between features from the two domains [14, 15], typically the Mean Maximum Discrepancy (MMD) [16] or some of its variants. However, minimizing the MMD can reduce feature-label correlation, which decreases class separability and can negatively impact downstream task performance.

Contrastive Learning (CL) is a learning paradigm, widely used in a Self-supervised Learning (SSL) setting, which consists in learning a representation where semantically similar features are close to one another and are distant to semantically different features. In Self-

supervised Contrastive Learning (SSCL), positive pairs are typically different views or transformed versions of the same input, while transformed inputs coming from different original inputs are considered as negative pairs. This training setting does not require expert annotations. SSCL is mostly used as a pretext task, that enables to exploit large unlabeled datasets, otherwise useless, for providing good weight initializations. The learnt representation is usually invariant to input transforms that are not present or are not especially relevant when training in the downstream task or for inference and evaluation, such as resizing, cropping, blurring, etc. [17]. In this context, SSCL has been widely applied in the medical imaging domain [18, 19], and for mammography images [20]. Supervised Contrastive Learning (SCL) is another variant of CL that uses annotations to assign positive and negative pairs: images with the same label are positive pairs, and images with different labels are negative pairs. SCL has been shown to improve robustness to corruptions in natural images, like noise, blur, and JPEG compression [21].

In this work, we seek to show the potential of SCL for Domain Adaptation, when training data consist in images with different post-processing. We argue that, by dragging features extracted by the DL-model with the same class to the same region of the feature space, SCL inherently performs Domain Adaptation and learns domain-invariant representations. In addition, SCL induces class-separability in the feature space, which improves classification performance. Supervised Contrastive Learning is compared to a training based on the Cross Entropy (CE) loss, a standard classification loss that does not perform Domain Adaptation.

2. MATERIALS AND METHODS

2.1. Classification model

Our ultimate goal was to develop a Convolutional Neural Network (CNN) model for whole 2D mammography images aiming to generate a per-image malignancy score, thresholded for binary cancer/no cancer classification. To achieve this, the process was divided into two stages.

First, a modified version of a previously proposed patch classifier [12] architecture was evaluated (Figure 1 – top). This involved training a DenseNet-121 to classify 512×512 pixels patches into five categories based on lesion type and pathology: normal, benign calcification, malignant calcification, benign mass, and malignant mass. DenseNet-121 was selected as backbone, as it achieved higher performance than other benchmarked CNNs [12]. The impact of the CNN architecture is out of the scope of this work. A Multi-layer Perceptron (MLP) projector was added to the model. This projector featured two hidden layers with 2048 units each, and an output layer of 1024 units. The projector was key to avoid perfect invariance, which can lower the classification effectiveness and it was used solely for the model training phase with Supervised Contrastive Learning (SCL) but not in the inference and evaluation stages. To initialize the patch classifier, two methods were explored: one using the ImageNet dataset and the other using the CBIS-DDSM dataset [22], a public scanned film mammography image collection

Next, the whole-image classifier was developed (Figure 1 – bottom). This classifier integrated the DenseNet-121 from the patch-classifier and a residual block was added. The whole-image classifier was trained on full mammograms. A similar MLP projector to that in the patch-classifier was added, consisting of one hidden layer

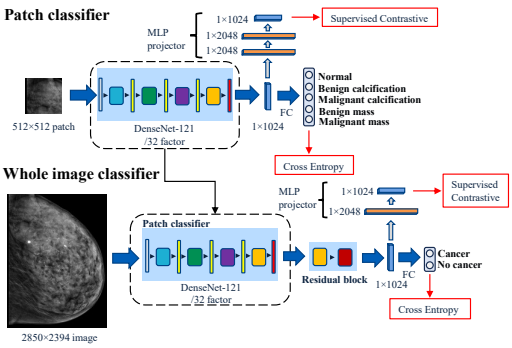


Fig. 1: Patch-classifier and whole image classifier architectures. FC: Fully Connected layer, MLP: Multi-layer Perceptron.

with 2048 units and an output layer of 1024 units. This projector was also employed only during training with SCL.

2.2. Datasets and post-processing

In this work, two distinct image domains, given by image post-processing differences, were considered. A sigmoid Look-Up Table (LUT) function, which is a contrast enhancement technique commonly used in mammography, was used for defining the two domains. However, the methodology developed in this work is applicable to any other image style or contrast transformation. Figure 2 shows an example image without LUT (domain 1) and with LUT applied (domain 2), and the pixel intensity histograms in logarithmic scale. Applying the LUT flattens the value of low-intensity and high-intensity pixels, but increases the intensity difference between pixels in the middle range, which contain the most useful information for diagnosis.

For training and evaluating the models, an internal GE HealthCare (GEHC) FFDM dataset was used. It contains 1539 cases, of which 363 are biopsy-proven cancers. The anonymized data were collected from a single institution in France following the EU General Data Protection Regulation. The dataset was split into training (1237 cases), validation (201 cases), and testing (101 cases) sets, using the method outlined in [12].

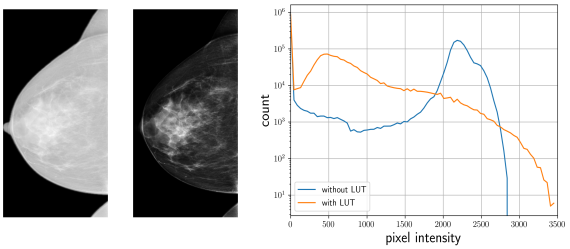
From the GEHC dataset, three distinct datasets were created for training and validation purposes:

- An *augmented dataset*, which includes two versions of each image: one with the LUT applied and the other without.
- *Mixed datasets*, formed by randomly dividing the GEHC dataset into two groups. One group had images with the LUT applied, while the other did not.
- *Mixed downsampled datasets*, generated by initially randomly excluding some images from the GEHC dataset using four keep ratios (0.6, 0.7, 0.8, and 0.9), followed by the creation of mixed datasets as described above.

Four mixed datasets and four mixed downsampled datasets for each keep ratio were constructed. The test set used for evaluation contained the two versions of each image (with and without LUT).

2.3. Contrastive-based Domain Adaptation

To learn a domain-invariant representation where classes are separated, the Supervised Contrastive loss was used [21]. This loss,



(a) Without LUT **(b)** With LUT **(c)** Pixel intensity histogram

Fig. 2: Illustration of a single FFDM image: (a) without LUT application, (b) with LUT application, and (c) the pixel intensity histogram in logarithmic scale for both post-processings.

calculating the similarity between pairs of features of the last convolutional layer with a temperature parameter τ, is given by:

$$\mathcal{L}_B = \sum_{i=1}^{B} \frac{-1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \log \frac{e^{z_i^T \cdot z_j/\tau}}{\sum_{l \in \mathcal{A}_i} e^{z_i^T \cdot z_l/\tau}}. \quad (1)$$

In Equation (1), B is the batch size, z_i is the i-th feature vector, $\mathcal{P}_i$ the set of all features that form positive pairs with z_i, and $\mathcal{A}_i$ the set of all the other features (positive and negatives), excluding feature z_i. Pairs of features of the same class are considered as positive, while pairs of features with different class are considered as negative. When the loss in Equation (1) is minimized, the features that correspond to images of the same class are pulled to the same region of the feature space, and the features corresponding to different classes are pushed to different regions of the feature space.

A model trained solely with the Cross Entropy (CE) loss, and thus without Domain Adaptation, was compared to a model trained using the Supervised Contrastive loss. As suggested in [21], the model was first trained with the Supervised Contrastive loss to extract domain-invariant features. Then, the feature extraction layers were frozen, and only the final linear classification layer was trained with CE. This model is denoted as SupContr, as the feature extraction was solely trained with the Supervised Contrastive loss. As a third training strategy, the SupContr model was fully re-trained using the Cross Entropy loss (without freezing the feature extraction layers), resulting in a model denoted as SupContr+CE. These three training strategies (CE, SupContr, and SupContr+CE) were compared for the patch-classifier and whole image classifier.

The CE, SupContr, and SupContr+CE models were trained on the four mixed datasets, and on the augmented dataset, for patch classification and whole image classification. The impact of the dataset size on the patch classifier was assessed, by training it on the mixed downsampled datasets and evaluating it on the same test set. In addition, the generalization capability of the whole image classifier was assessed on InBreast [23], a publicly available dataset of 2D mammography images. As InBreast does not contain biopsy-proven cancer labels, our model was fine-tuned for BIRADS 4 and 5 (i.e. suspicious and highly suggestive for malignancy) vs. BIRADS 1, 2, and 3 (i.e. negative, benign, probably benign) binary classification. For this, the InBreast dataset was split into training (288 cases), validation (46 cases) and testing (75 cases) sets, with the same stratification strategy used for the GEHC dataset. To keep the learnt representation fixed, the feature extractor was frozen during InBreast fine-tuning, and only the output linear layer was updated.

	CE	SupContr	SupContr + CE
mixed dataset 1	0.737 ± 0.008 (n.a.)	0.836 ± 0.006 (< 0.001)	$\mathbf{0.840 \pm 0.006}$ ($\mathbf{< 0.001}$)
mixed dataset 2	0.692 ± 0.008 (n.a.)	0.820 ± 0.007 (< 0.001)	$\mathbf{0.833 \pm 0.006}$ ($\mathbf{< 0.001}$)
mixed dataset 3	0.750 ± 0.008 (n.a.)	$\mathbf{0.850 \pm 0.006}$ ($\mathbf{< 0.001}$)	0.842 ± 0.006 (< 0.001)
mixed dataset 4	0.745 ± 0.007 (n.a.)	$\mathbf{0.846 \pm 0.005}$ ($\mathbf{< 0.001}$)	$\mathbf{0.846 \pm 0.006}$ ($\mathbf{< 0.001}$)
augmented dataset	0.871 ± 0.006 (n.a.)	0.878 ± 0.005 (< 0.001)	$\mathbf{0.887 \pm 0.005}$ ($\mathbf{< 0.001}$)

Table 1: Mean one-vs-one AUC, 95% CI, and p-value of the patch-classifier, with ImageNet weight initialization.

The patch-classifiers were evaluated and compared in terms of the mean one-vs-one AUC, a standard multi-class classification metric, and the whole image classifiers in terms of the AUC. Bootstrapping was used for calculating the 95% Confidence Intervals (CI) and the p-values with respect to the CE models, using the Welch's t-test. In addition, t-SNE [24] was used for visualizing the extracted features.

3. RESULTS

3.1. Patch-classifier

Table 1 shows the results of the patch-classifier with ImageNet initialization. The best performing models of each dataset are noted in bold. From Table 1, SupContr and SupContr+CE outperform the CE for training on all the mixed datasets and on the augmented dataset. Figure 4a shows the t-SNE plot of the extracted features for the three models. While for the CE model features are more separated by domain than by class, the SupContr and SupContr+CE models are domain-invariant.

Table 2 shows the results of the patch-classifier with weights initialized from CBIS-DDSM. It can be seen that the AUC is generally higher than in Table 1, as the model was pre-trained on mammography images. In contrast to the results of Table 1, SupContr and SupContr+CE fail to outperform CE in all the datasets. In Figure 4b, which shows the features t-SNE plot, it can be seen that the CE model is more feature invariant when pre-trained with CBIS-DDSM than with ImageNet, which decreases the impact of Domain Adaptation on AUC. However, the features of normal patches are not adapted in the CE model, while they are adapted in the SupContr and SupContr+CE models. SupContr+CE trades-off an invariant representation with a slight or no decrease in AUC, depending on the dataset. The patch-classifiers pre-trained on CBIS-DDSM were used for extension to the whole image classifiers, as they all outperform the ones initialized from ImageNet.

Figure 3 shows the mean one-vs-one AUC of the patch-classifier with CBIS-DDSM initialization trained on 4 mixed downsampled datasets for each keep ratio, and on the 4 mixed datasets (i.e., keep ratio = 1.0). It can be seen that SupContr+CE outperforms CE (and SupContr) when the dataset size decreases, despite failing to outperform CE in the mixed datasets (keep ratio = 1.0). This dataset size effects shows that Domain Adaptation increases AUC when training with smaller datasets, even when weights are initialized from another mammography image dataset.

Similar results to the ones observed in Tables 1 and 2 were observed in terms of the accuracy and mean one-vs-rest AUC. The only exception is in the augmented dataset, in which SupContr+CE fails to outperform CE in terms of one-vs-rest AUC for ImageNet initialization. This can be explained by the dataset size effect observed in Figure 3, as the augmented dataset contains twice as many images as the mixed datasets. A detailed comparison of the results on these two metrics is out of the scope of this work.

	CE	SupContr	SupContr + CE
mixed dataset 1	0.905 ± 0.005 (n.a.)	0.898 ± 0.005 (< 0.001)	$\mathbf{0.920 \pm 0.005}$ ($\mathbf{< 0.001}$)
mixed dataset 2	$\mathbf{0.922 \pm 0.004}$ **(n.a.)**	0.889 ± 0.005 (< 0.001)	0.910 ± 0.005 (< 0.001)
mixed dataset 3	$\mathbf{0.926 \pm 0.004}$ **(n.a.)**	0.876 ± 0.005 (< 0.001)	0.922 ± 0.004 (< 0.001)
mixed dataset 4	$\mathbf{0.919 \pm 0.004}$ (n.a.)	0.867 ± 0.005 (< 0.001)	$\mathbf{0.920 \pm 0.004}$ **(n.s.)**
augmented dataset	$\mathbf{0.927 \pm 0.004}$ **(n.a.)**	0.880 ± 0.005 (< 0.001)	0.919 ± 0.004 (< 0.001)

Table 2: Mean one-vs-one AUC, 95% CI, and p-value of the patch-classifier, with CBIS-DDSM weight initialization.

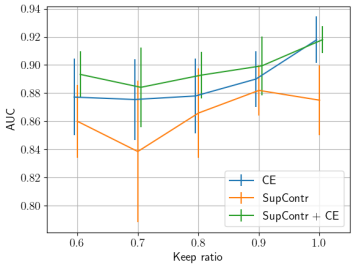
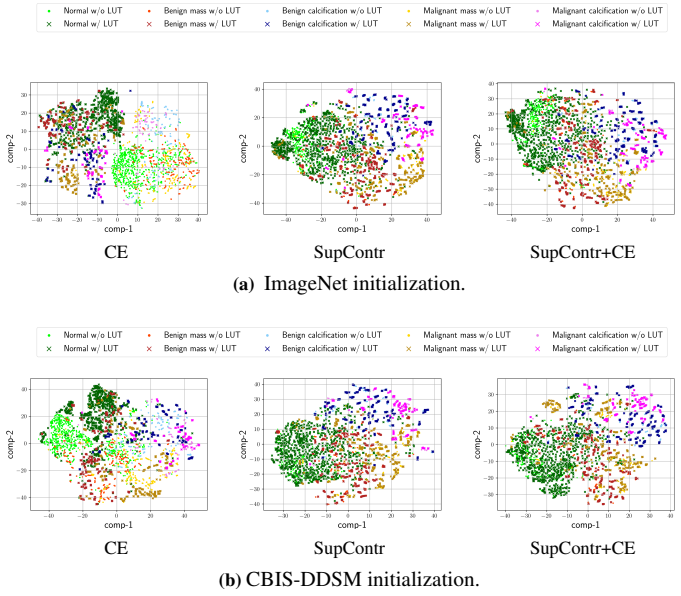


Fig. 3: Mean one-vs-one AUC with 95% CI in the test set, of the models trained on the 4 mixed downsampled datasets, by keep ratio.



(a) ImageNet initialization.

(b) CBIS-DDSM initialization.

Fig. 4: t-SNE plots of the features from the patch-classifier, indicating class and domain.

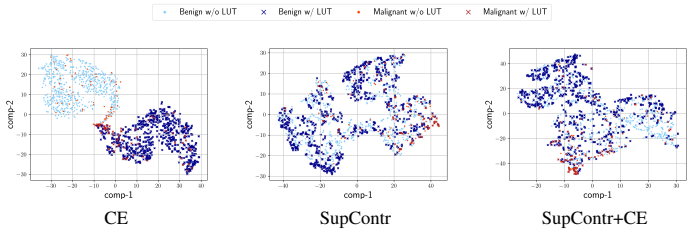


Fig. 5: t-SNE plot of the features from the whole image classifier, indicating class and domain.

3.2. Whole image classifier

Table 3 shows the AUC in the test set for the CE, SupContr, and SupContr+CE models for the whole image classifier. Each model used the corresponding patch-classifier of Section 3.1, with CBIS-DDSM initialization. In Table 3, the SupContr and SupContr+CE models significantly outperform the CE model in all the considered datasets. In the augmented dataset, SupContr+CE reaches an AUC of 0.816 ± 0.042, a 10% increase with respect to the CE model. SupContr+CE outperforms SupContr in all the datasets, except for the mixed dataset 3, in which the AUC between the two models is not statistically significant. Furthermore, the contrastive-based models trained on all the mixed datasets outperform the CE model trained on the augmented dataset, despite the latter having been trained on twice the amount of data. In addition, the SupContr+CE model exhibits superior generalization by achieving a 13% AUC increase in InBreast, with the value rising from 0.733 ± 0.096 to 0.831 ± 0.071. The higher CI in InBreast with respect to the GEHC dataset can be explained by the size of the two datasets (409 and 1539 cases, respectively). In the t-SNE plot of Figure 5 it can be seen that the features of the CE model can be easily separated by domain, despite the features of the CE patch-classifier being domain-invariant for most classes. We hypothesize that this is caused by the maladaptation of the normal patches for the CE model in Figure 4b, as every mammography image contains many normal regions. On the other hand, the features of the SupContr and SupContr+CE models are domain-invariant.

	CE	SupContr	SupContr + CE
mixed dataset 1	0.723 ± 0.056 (n.a.)	0.772 ± 0.053 (< 0.001)	$\mathbf{0.793 \pm 0.049}$ ($\mathbf{< 0.001}$)
mixed dataset 2	0.747 ± 0.046 (n.a.)	0.801 ± 0.041 (< 0.001)	$\mathbf{0.811 \pm 0.047}$ ($\mathbf{< 0.001}$)
mixed dataset 3	0.716 ± 0.062 (n.a.)	$\mathbf{0.786 \pm 0.049}$ ($\mathbf{< 0.001}$)	$\mathbf{0.781 \pm 0.054}$ ($\mathbf{< 0.001}$)
mixed dataset 4	0.714 ± 0.061 (n.a.)	0.762 ± 0.051 (< 0.001)	$\mathbf{0.794 \pm 0.048}$ ($\mathbf{< 0.001}$)
augmented dataset	0.745 ± 0.050 (n.a.)	0.763 ± 0.058 (< 0.001)	$\mathbf{0.816 \pm 0.042}$ ($\mathbf{< 0.001}$)
InBreast	0.733 ± 0.096 (n.a.)	0.746 ± 0.083 (n.s.)	$\mathbf{0.831 \pm 0.071}$ ($\mathbf{< 0.001}$)

Table 3: AUC, 95% CI, and p-value of the whole image classifier

4. CONCLUSION

This work shows the potential of Contrastive Learning as a Domain Adaptation technique, that solves the hidden covariate shift. It significantly increased AUC for whole mammography image classification, and showed higher generalization in InBreast. For the patch-classifier, DA increased AUC when training with smaller sub-sets of the GEHC dataset, or when initializing weights from natural images. This suggests that the benefits of DA might increase when less training data is available, or when the model does not benefit from Transfer Learning from a similar dataset and task. The Contrastive-based DA methodology developed in this work does not depend on the LUT post-processing used for the experiments, and its applicability can be extended to other modalities and beyond medical imaging.

However, as our approach is based on the Supervised Contrastive loss, it relies on lesion-level (for sampling the patches from the mammography images) and image-level annotations. This introduces an important challenge due to the required workload in obtaining accurate annotations from clinicians and radiologists, which in turn increases the economic cost. A potential solution to mitigate this challenge is to adopt Self-supervised variants of the Contrastive loss, such as the NT-Xent loss [25], which could reduce the reliance on extensive manual annotations.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study develops an AI algorithm using an anonymized database for which no ethical approval was required.

6. ACKNOWLEDGMENTS

The authors wish to acknowledge the French ANRT (Association Nationale de la Recherche et de la Technologie) for partially funding this research, under CIFRE grant number 2021/0422. The authors also wish to acknowledge the creators and contributors of the Monai project [26], which was used in this work.

7. REFERENCES

[1] Antonio Torralba and Alexei A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, 2011, pp. 1521–1528.

[2] Ishaan Gulrajani and David Lopez-Paz, "In search of lost domain generalization," *CoRR*, vol. abs/2007.01434, 2020.

[3] Lidia Garrucho, Kaisar Kushibar, Socayna Jouide, Oliver Diaz, Laura Igual, and Karim Lekadir, "Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study," *Artificial Intelligence in Medicine*, vol. 132, pp. 102386, 2022.

[4] Antoine de Mathelin, Francois Deheeger, Mathilde Mougeot, and Nicolas Vayatis, *From Theoretical to Practical Transfer Learning: The ADAPT Library*, pp. 283–306, Springer International Publishing, Cham, 2023.

[5] Meelis Kull and Peter A. Flach, "Patterns of dataset shift," 2014, First International Workshop on Learning over Multiple Contexts (LMCE) at ECML-PKDD.

[6] Amos Storkey, J Quiñonero-Candela, M Sugiyama, A Schwaighofer, and ND Lawrence, *When Training and Test Sets Are Different: Characterizing Learning Transfer*, pp. 3–28, Neural Information Processing Series. Yale University Press in association with the Museum of London, Dec. 2008.

[7] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

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

[9] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," 2016.

[10] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng, "Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss," *CoRR*, vol. abs/1804.10916, 2018.

[11] Amelia Jiménez-Sánchez, Mickael Tardy, Miguel Ángel González Ballester, Diana Mateus, and Gemma Piella, "Memory-aware curriculum federated learning for breast cancer classification," *CoRR*, vol. abs/2107.02504, 2021.

[12] Gonzalo Iñaki Quintana, Zhijin Li, Laurence Vancamberg, Mathilde Mougeot, Agnès Desolneux, and Serge Muller, "Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification," *Bioengineering*, vol. 10, no. 5, 2023.

[13] Wouter M. Kouw and Marco Loog, "A review of domain adaptation without target labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 766–785, mar 2021.

[14] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[15] Mahsa Baktashmotlagh, Mehrtash Har, i, and Mathieu Salzmann, "Distribution-matching embedding for visual domain adaptation," *Journal of Machine Learning Research*, vol. 17, no. 108, pp. 1–30, 2016.

[16] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola, "A kernel method for the two-sample-problem," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. 2006, vol. 19, MIT Press.

[17] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum, "A cookbook of self-supervised learning," 2023.

[18] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *CoRR*, vol. abs/2006.10511, 2020.

[19] Nanqing Dong and Irina Voiculescu, "Federated contrastive learning for decentralized unlabeled medical images," 2021.

[20] Zhenjie Cao, Zhicheng Yang, Yuxing Tang, Yanbo Zhang, Mei Han, Jing Xiao, Jie Ma, and Peng Chang, "Supervised contrastive pre-training for mammographic triage screening models," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, Eds., Cham, 2021, pp. 129–139, Springer International Publishing.

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *CoRR*, vol. abs/2004.11362, 2020.

[22] Rebecca Lee, Francisco Gimenez, Assaf Hoogi, Kanae Miyake, Mia Gorovoy, and Daniel Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, pp. 170177, 12 2017.

[23] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso, "Inbreast: Toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[24] Laurens van der Maaten and Geoffrey Hinton, "Viualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020.

[26] Project MONAI. Zenodo, "The MONAI Consortium," 2022.