Sharing Generative Models Instead of Private Data: A Simulation Study on Mammography Patch Classification

Zuzanna Szafranowska*[a], Richard Osuala*[a], Bennet Breier[a], Kaisar Kushibar[a], Karim Lekadir[a], and Oliver Diaz[a]

[a]Barcelona Artificial Intelligence in Medicine Lab, Faculty of Mathematics and Computer Science, University of Barcelona, Spain

ABSTRACT

Early detection of breast cancer in mammography screening via deep-learning based computer-aided detection systems shows promising potential in improving the curability and mortality rates of breast cancer. However, many clinical centres are restricted in the amount and heterogeneity of available data to train such models to (i) achieve promising performance and to (ii) generalise well across acquisition protocols and domains. As sharing data between centres is restricted due to patient privacy concerns, we propose a potential solution: sharing trained generative models between centres as substitute for real patient data. In this work, we use three well known mammography datasets to simulate three different centres, where one centre receives the trained generator of Generative Adversarial Networks (GANs) from the two remaining centres in order to augment the size and heterogeneity of its training dataset. We evaluate the utility of this approach on mammography patch classification on the test set of the GAN-receiving centre using two different classification models, (a) a convolutional neural network and (b) a transformer neural network. Our experiments demonstrate that shared GANs notably increase the performance of both transformer and convolutional classification models and highlight this approach as a viable alternative to inter-centre data sharing. Find our code at https://github.com/zuzaanto/mammo_gans_iwbi2022

Keywords: Generative Adversarial Networks, Data Sharing, Patient Privacy, Mammography, Synthetic Data, Breast Cancer, Transformer, Deep Learning, Swin Transformer, Computer-Aided Detection

1. INTRODUCTION

With an estimated worldwide incidence rate of 47.8 per 100,000 people in 2020 (both sexes, all ages), breast cancer is the cancer type with the highest prevalence in the world. It accounts for an estimated 2.22 million new cases and 684,996 deaths each year.[1] Screening mammography (MMG) provides early detection and contributes to reducing breast cancer mortality.[2] However, MMG images are limited by their error rates due to tissue superposition which could lead to underdiagnosis of significant breast cancers (false negatives) and overdiagnosis of insignificantly abnormal or healthy cases (false positives).[2,3] In this regard, deep learning based computer-aided detection (CADe) systems have shown great promise in improving and automating the decision making process of mammograms.[2,4] However, deep learning methods are known to require large amounts of training data to achieve accurate, reliable and robust performance.

Scarcity of expert-annotated medical images often constrains deep learning based methods to be trained and evaluated on a small dataset coming from a single centre.[5] Accordingly, such methods suffer from lack of generalisation and robustness.[6] A solution to this problem is inter-centre data sharing. However, clinical centres are constrained from sharing sensitive patient data due to technical, legal, and most importantly, ethical concerns.[3,7]

Synthetic images generated by Generative Adversarial Networks (GANs)[8] have been shown to be of high perceived visual realism for mammography and to improve downstream tasks including cancer detection, tumour

*equal contribution
Z.S.: E-mail: z.szafranowska@gmail.com
R.O.: E-mail: richard.osuala@gmail.com

16th International Workshop on Breast Imaging (IWBI2022), edited by Hilde Bosmans,
Nicholas Marshall, Chantal Van Ongeval, Proc. of SPIE Vol. 12286, 122860Q
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2625781

segmentation and classification.[3] Therefore, in this work, we hypothesize that GANs can overcome the inter-centre data sharing constraints. After learning the real data distribution, GANs can generate synthetic data with limited risk of revealing sensitive patient information.[3,9] Then, a clinical centre can share a trained generative model that will act as a proxy for the real patient data.

In this work, we investigate the application of GANs as substitutes for multi-centre real patient data. Using three well known mammography datasets acquired at different sites from Portugal[10,11] and UK,[12] we simulate three different centres, where one centre receives GANs from the two remaining centres to augment the size and heterogeneity of its training dataset. Figure 1 shows our simulated privacy-preserving data sharing setup using GANs. We evaluate the performance of a convolutional neural network (CNN)[13,14] and a transformer neural network[15] for healthy versus non-healthy tissue classification in a region of interest (ROI) using two training strategies: 1) using data from only a single centre; and 2) using additional synthetic data from other centres. We demonstrate through our experiments that augmenting single centre data using GAN generated image ROIs considerably improves classification performance.

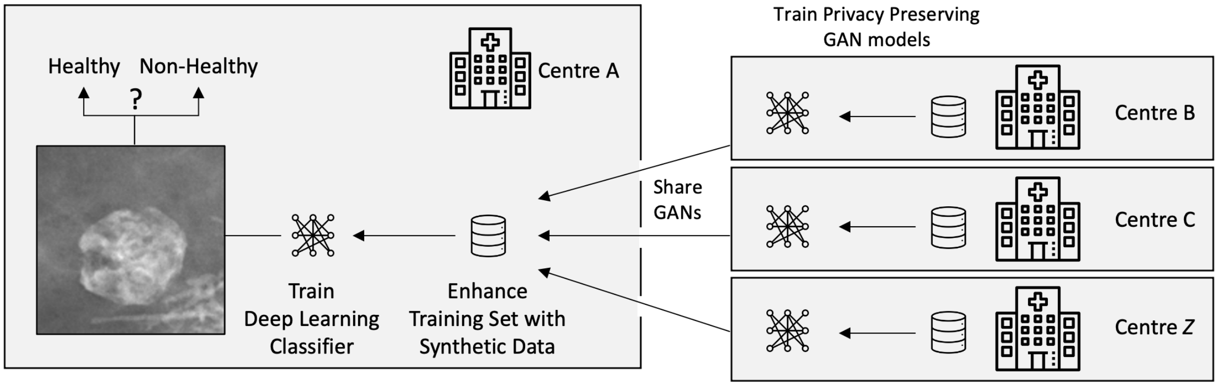

Figure 1. Overview of sharing generative models (e.g., GANs) as substitutes for multi-centre patient data to (i) overcome data-scarcity in single centres, and (ii) to enhance prediction performance and robustness. The latter is exemplified on the task of classifying the presence of a lesion in a mammography patch.

2. METHODOLOGY

2.1 Experimental setup

Our methodological framework consists of two parts. First, we utilise a GANs to learn the distribution of the data from data-providing centres (Centres B, C, etc. in Figure 1). These trained GANs from each centre are then sent to the main centre (Centre A in Figure 1) for further use. Second, the main centre trains a classification model that classifies whether a lesion is present in MMG patches.

The training data for the classification model in the main centre is augmented by GAN-generated lesion patches trained on data from other centres. To maintain class balance between *healthy* and *non-healthy* (i.e., has lesion) labels, we add the same number of healthy patches extracted from existing healthy control MMGs from centre A. Finally, the performance of the classification model trained only on data from centre A is compared to the performance of the same classification model trained on the combination of centre A data and synthetic data from other centres.

In our experiments, we use INbreast[10] and BCDR[11] mammography datasets and an additional pretrained GAN trained on the OPTIMAM dataset.[12] INbreast consists of digital MMGs from 115 patients, totalling 410 images. A total number of lesions is 3028 (including 116 masses), which corresponds to the number of INbreast patches in our experiments that include both benign and malignant masses and calcifications. The BCDR dataset consists of both digital and film-scanned MMG images from 1010 patients, totalling 1493 lesions (including 639 masses). The pretrained GAN was trained on 2215 masses extracted from the OPTIMAM dataset.

In particular, we gathered experimental results for testing on data from centre A (represented by the INbreast dataset) after training on data from (i) centre A (INbreast), (ii) centre A (INbreast) and centre B (BCDR), (iii) centre A (INbreast) and synthetic data from centre B (BCDR) and/or centre C (OPTIMAM). We conducted these experiments (i to iii) for two scenarios. First, the scenario where 100% of the centre A (INbreast) training data is used for training. And secondly, the setting where centre A is experiencing data scarcity with its internal training data set randomly reduced by 50%. Also we separately conduct each experiment for tumour masses specifically, and for breast lesions (calcifications, masses, etc.) in general. As shared generative model of centre B, we train both a Deep Convolutional Generative Adversarial Network (DCGAN)[16] and a Wasserstein GAN with Gradient Penalty (WGAN-GP),[17] while for centre C we reuse a pretrained DCGAN from Alyafi et al[18,19] via the *medigan* model sharing library[20†]. We train, evaluate, and report results for two different classification models in centre A, namely (a) a convolutional neural network (CNN) and (b) a Swin transformer neural network.[21]

The classification performance is evaluated using the accuracy, F1-score, area under receiver operating characteristic curve (AUROC), and area under precision-recall curve (AUPRC) metrics. To further increase the informative value of our results, we run all experiments three times with a different random seed in each run and report the resulting mean and standard deviation of each metric.

All classification experiments were run on a machine equipped with NVIDIA GTX 1070 8GB GPU, using the PyTorch library.[22] Our GANs were trained on a NVIDIA RTX 2080 Super 8GB GPU, also using PyTorch.

2.2 Mammogram Patch Extraction

Healthy patches are extracted from INbreast MMGs of healthy breasts that have no annotation that indicates the presence of a lesion. We generate a number of bounding boxes randomly defined within an entirely healthy breast image, ensuring that these patches never contain more than 40% of background pixels.

Non-healthy patches are crops containing both malignant and benign lesions of any breast imaging-reporting and data system (BI-RADS) score and of any lesion type present in the datasets including masses, calcifications, microcalcifications, and architectural distortions. We use the lesions contour information specified in the original datasets[10,11] and create a bounding box enclosing it. Then, we create a square patch around that bounding box, ensuring a margin of 60 pixel in each direction from the lesion bounding box. If the margin extends beyond the mammogram's border, a translation is performed to the patch to ensure it is fully within the mammogram's limits.

After specifying the patch bounding boxes, we use the same pre-processing routine at training time for both healthy and non-healthy patches to ensure that class-specific pre-processing artefacts are not introduced that would otherwise be easily distinguishable by the classifier. Each patch, defined by its bounding boxes, is first zoomed and then translated by normally random amounts. In doing so, we offset the patch from its original bounding box, which would otherwise always be close to the centre of the patch, and as a strategy of data augmentation. Finally, each patch is resized to 128x128 pixels using inter-area interpolation, whereby image ratios are maintained.

2.3 Generative Adversarial Networks as Data Substitute

GANs[8] are a type of generative model and are composed of the discriminator (D) and the generator network (G) that compete against each other in a two-player zero-sum game defined by the value function shown in equation 1.

$$\min_{G}\max_{D} V(D,G) = \min_{G}\max_{D}[\mathbb{E}_{x\sim\mathbb{P}_{data}}[log(D(x))] + \mathbb{E}_{z\sim\mathbb{P}_{z}}[log(1-D(G(z)))]] \tag{1}$$

2.3.1 DCGAN Generator Trained on BCDR Data

We adopt a DCGAN[16] with some adjustments suggested in[18] such as one-sided label smoothing in range [0.8, 1.1], and a discriminator with a kernel size 6 instead of 4. We train our DCGAN to learn the distribution of the training data consisting of 128x128 pixel grayscale mammogram patches. The DCGAN learns a mapping between a vector containing 100 numerical values to a mammogram patch containing a breast lesion. The GAN training data is augmented by random (p=0.5) horizontal and (p=0.5) vertical flipping and uses a batch size

[†] https://medigan.readthedocs.io/

of 16. Depending on the classification objective of our experiments the GAN is either trained on patches of all lesion types, or specifically on patches containing a mass.

We train our DCGAN on the BCDR dataset for 3000 epochs to generate synthetic patches similar to the real patches. After each epoch, we visually assessed the fidelity of the mammogram patches generated from a set of fixed noise vectors. During training we noticed an increase in fidelity, but also a decrease in diversity of the generated patches: The similarity of some of the generated lesions indicated the occurrence of mode collapse, the state where the generator has learned to repeatedly generate a limited subset of samples to fool the discriminator. Interestingly, the same fixed noise vector input created varying lesion shapes and textures in different training iterations indicating a high diversity across iterations and epochs.

Observing this behaviour two measures were taken. Firstly, to maximise diversity, we store the weights of our DCGAN during training on each 50th epoch starting in epoch 500. After training, we generate mammogram patches using the stored weights (epoch 500 to 3000) to generate the synthetic dataset. Secondly, we explore further GAN alternatives less prone to mode collapse and, based on our DCGAN network architecture, implement Wasserstein GAN with Gradient Penalty (WGAN-GP).[17]

2.3.2 WGAN-GP Generator Trained on BCDR Data

To overcome mode collapse in DCGAN and to increase training stability, we follow the approach of[23] of substituting DCGAN's binary cross-entropy loss with a Wasserstein distance based loss function. We apply the Wasserstein with gradient penalty[17] loss function to the setup described in 2.3.1. Equation 2 displays the WGAN-GP loss function with penalty coefficient λ (set to 10 in our experiments) and distribution $\mathbb{P}_{\hat{x}}$ sampling uniformly along straight lines between pairs of points from the generator distribution $\mathbb{P}_g$, and the data distribution $\mathbb{P}_{data}$.

$$L \;=\; E_{\tilde{x}\sim\mathbb{P}_g}\left[D\left(\tilde{x}\right)\right] \;-\; E_{x\sim\mathbb{P}_{data}}\left[D\left(x\right)\right] \;+\; \lambda\; E_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{x}}D\left(\hat{x}\right)\right\|_2 - 1\right)^2\right] \tag{2}$$

The discriminator (alias critic) is updated 5 times per generator update. Further, we remove the batch normalization layers from the DCGAN discriminator, as WGAN-GP penalizes the norm of the discriminator's gradient per input sample rather than per batch. Both gradient penalty (WGAN-GP) [17] and weight clipping (WGAN)[24] enforce a 1-Lipschitz constraint while the former additionally avoids model capacity underuse, which motivates our choice of WGAN-GP. We train WGAN-GP for 10000 epochs on mammogram patches that contain masses and for 2700 epochs on mammogram patches that contain any type of lesion. In both cases, due to high image diversity upon visual assessment, only the checkpoint from the last epoch was used to generate the synthetic images for our subsequent classification experiments.

2.3.3 DCGAN Generator Pretrained on OPTIMAM Data

Furthermore, we use the DCGAN published by Alyafi et al[18, 19] from the *medigan*[20] library pre-trained on 2215 mass patches from the OPTIMAM dataset to generate an additional dataset of synthetic mammogram patches containing masses. Figure 2 illustrates manually selected synthetic images generated by Alyafi et al's DCGAN, and our BCDR-trained DCGAN (2.3.1) and WGAN-GP (2.3.2). As opposed to our GANs, Alyafi et al's DCGAN only generates patches containing a mass and is not trained on other types of lesions.

2.4 Classification Models

All experiments with the classifiers are validated and tested on INbreast[10] data that have been split previously into training (approx. 70%), validation (approx. 15%) and testing (approx. 15%) subsets. We ensure that (1) all images from a single patient are always in only one of the subsets, (2) healthy and non-healthy patches are always present in similar amounts, and (3) the distributions of breast density values in each of the subsets are as similar as possible. Dataset splits are performed separately, once for the experiments where we classify any type of lesion (masses, calcifications, etc) and once for the experiments where we classify exclusively the lesion type "mass". In the cases when the patches with lesions stem from two GANs, we add half the patches from each GAN. Additionally, for all types of augmentations, we add as many healthy patches from INbreast as needed to maintain the class balance between healthy and non-healthy training data. Overall, we extract

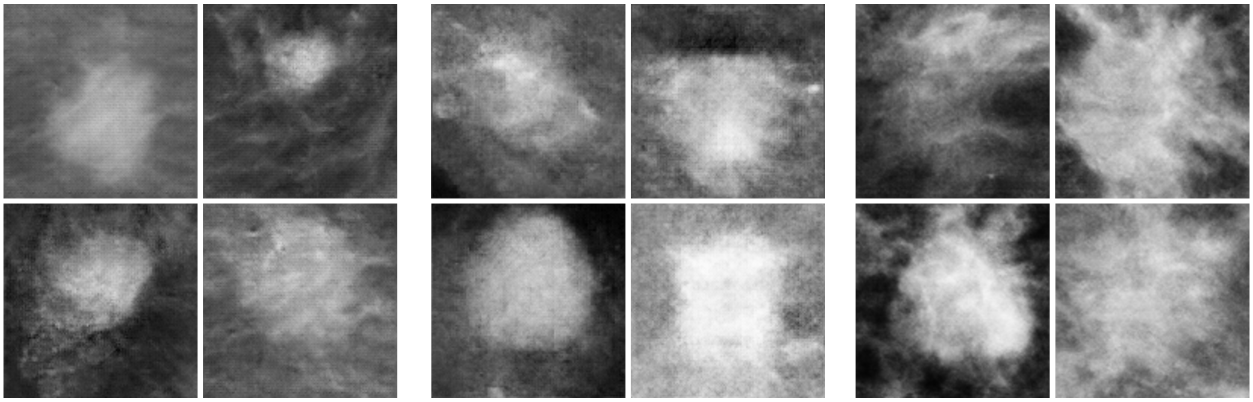

Figure 2. Left column: synthetic patches generated from BCDR (using our DCGAN). Middle column: synthetic patches generated from BCDR (using our WGAN-GP). Right column: synthetic patches generated from OPTIMAM (using the pretrained DCGAN of Alyafi et al[18,19]).

1374 (70 in mass classification) real patches for each class from INbreast for training. Then, 1200 additional synthetic patches, either all lesions or only masses and either from one or from two centres are added to the training as augmentation. In cases of augmentation with real BCDR data we extract patches from both digital and film-scanned BCDR mammograms and add 1148 patches for lesion classification and 500 patches for mass classification to the class "non-healthy".

2.4.1 Convolutional Classification Model

As a baseline classification model, we implement a CNN, consisting of four convolutional followed by two fully-connected layers, with ReLU[25] activation functions. After each convolutional layer and after the first fully-connected layer, batch normalization[26] is performed. Also, dropout[27] is applied between the two fully-connected layers to introduce regularization during training and reduce the risk of overfitting. Finally, the network's output is passed through a softmax function followed by a logarithm. Our loss function is defined as cross-entropy. For training the network, we first initialize the weights randomly after setting a random seed. Then, we use stochastic gradient descent optimization of the loss function with the learning rate of 0.001 and momentum of 0.9. To obtain the best model, after each epoch, we perform validation on a separate subset of data, and only save the model if it achieves superior AUPRC score as compared to the saved models from previous epochs. The upper limit of epochs is set to 100.

2.4.2 Swin Transformer Classification Model

For further corroboration of our results, we run the same experiments using Swin Transformer[21] as classification model. This increases the generalisability of approach by comparing two considerably different methodological frameworks that showed state-of-the-art performances in vision tasks. Due to heavy exploitation of self-attention mechanisms, Transformers consider relations between all pair-wise local regions in the image. Accordingly, they do not assume that related features are close to each other in the image and eliminate such inductive bias that is present in CNNs.

Swin Transformers are an extension of hierarchical vision transformers, which implements shifted windows mechanism as a way to limit self-attention computation to non-overlapping local windows, while allowing for cross-window connection. In our experiments, we used the original Swin Transformer setup as in Liu et al.[21] Therefore, we resized all the input patches by resampling with respect to pixel area relation to 224 × 224 and stacked them to obtain a three channel input.

3. RESULTS

Tables 1 and 2 summarise our experimental results. Each experiment yields insights into different combinations of datasets, along three dimensions: First, we use different data augmentations for the training set, namely synthetic patches only from BCDR generated either with a WGAN-GP or a DCGAN, synthetic patches only from OPTIMAM (DCGAN), synthetic patches from both BCDR and OPTIMAM where for BCDR we used either a DCGAN or a WGAN-GP, or real patches from BCDR. Including real patches from BCDR serves as an expected upper bound for the experiments with synthetic patches. Second, we train the classifier once with the entire INbreast training set and once with only 50% thereof. Third, we either include all lesion-types in the non-healthy class or masses only.

Table 1. Results for classification of **all lesion types** including masses, calcifications, etc, using CNN (top) and Swin Transformer (bottom). The left-most column refers to the source of data added to the INbreast training set. "Both (a)" means augmenting with synthetic patches from both BCDR (WGAN-GP) and OPTIMAM (DCGAN), while "Both (b)" means BCDR (DCGAN) and OPTIMAM (DCGAN). All classifiers are tested on the INbreast testset. The best results per column are presented in bold font. Results are shown as mean(std).

		100% of the INbreast training data				50% of the INbreast training data			
	Augmentation	Accuracy	F1	AUROC	AUPRC	Accuracy	F1	AUROC	AUPRC
CNN	None	0.942(.005)	0.693(.020)	0.955(.007)	0.880(.014)	0.933(.010)	0.661(.035)	0.960(.004)	0.878(.013)
	BCDR (WGAN-GP)	0.938(.005)	0.673(.012)	0.957(.004)	0.839(.047)	0.929(.048)	0.669(.130)	0.953(.003)	0.863(.015)
	BCDR (DCGAN)	0.936(.030)	0.685(.099)	**0.966(.001)**	**0.887(.006)**	0.926(.034)	0.652(.110)	0.960(.005)	0.874(.006)
	OPTIMAM (DCGAN)	**0.945(.005)**	0.701(.014)	0.961(.007)	0.871(.010)	0.945(.002)	0.703(.004)	0.963(.003)	0.879(.007)
	Both (a)	0.943(.010)	0.701(.039)	0.965(.004)	0.863(.014)	**0.947(.020)**	**0.718(.075)**	**0.965(.005)**	**0.886(.014)**
	Both (b)	**0.945(.017)**	**0.704(.065)**	0.959(.005)	0.871(.013)	0.945(.013)	0.706(.044)	0.964(.007)	**0.886(.013)**
	Real BCDR	0.924(.001)	0.633(.001)	0.958(.001)	0.851(.003)	0.938(.007)	0.680(.028)	**0.965(.002)**	0.884(.016)
Swin Transformer	None	0.914(.065)	0.634(.161)	0.951(.004)	0.859(.010)	0.949(.040)	0.734(.138)	0.933(.004)	0.860(.006)
	BCDR (WGAN-GP)	0.912(.082)	0.647(.186)	0.952(.001)	0.864(.002)	**0.974(.005)**	0.824(.027)	0.937(.015)	0.860(.006)
	BCDR (DCGAN)	**0.974(.002)**	**0.826(.009)**	0.955(.004)	0.874(.005)	0.959(.030)	**0.880(.066)**	**0.958(.007)**	**0.916(.067)**
	OPTIMAM (DCGAN)	0.946(.038)	0.724(.120)	0.957(.002)	0.870(.002)	0.927(.043)	0.667(.150)	0.953(.016)	0.871(.029)
	Both (a)	0.922(.040)	0.637(.112)	0.940(.019)	0.852(.022)	0.940(.029)	0.692(.104)	0.957(.006)	0.872(.010)
	Both (b)	0.969(.002)	0.804(.008)	0.957(.004)	0.876(.009)	0.965(.011)	0.783(.050)	0.956(.002)	0.877(.008)
	Real BCDR	0.942(.033)	0.713(.132)	**0.962(.003)**	**0.897(.001)**	0.854(.001)	0.500(.002)	0.952(.002)	0.873(.001)

Table 2. Results for classification of **tumour masses** using CNN (top) and Swin Transformer (bottom). The left-most column refers to the source of data added to the INbreast training set. "Both (a)" means augmenting with synthetic patches from both BCDR (WGAN-GP) and OPTIMAM (DCGAN), while "Both (b)" means BCDR (DCGAN) and OPTIMAM (DCGAN). All classifiers are tested on the INbreast testset. The best results per column are presented in bold font. Results are shown as mean(std).

		100% of the INbreast training data				50% of the INbreast training data			
	Augmentation	Accuracy	F1	AUROC	AUPRC	Accuracy	F1	AUROC	AUPRC
CNN	None	0.895(.019)	0.935(.015)	0.962(.010)	0.992(.002)	0.897(.013)	0.938(.009)	0.955(.016)	0.991(.003)
	BCDR (WGAN-GP)	**0.943(.023)**	**0.966(.014)**	0.973(.010)	**0.995(.002)**	0.872(.013)	0.919(.008)	0.962(.009)	0.992(.002)
	BCDR (DCGAN)	0.910(.016)	0.944(.010)	**0.976(.005)**	**0.995(.001)**	0.904(.020)	0.941(.014)	**0.968(.014)**	**0.993(.003)**
	OPTIMAM (DCGAN)	0.876(.072)	0.919(.051)	0.971(.010)	0.994(.002)	0.925(.033)	0.954(.022)	**0.968(.008)**	**0.993(.002)**
	Both (a)	0.908(.062)	0.942(.043)	0.970(.007)	0.994(.002)	**0.933(.020)**	**0.960(.012)**	0.946(.026)	0.983(.011)
	Both (b)	0.925(.019)	0.954(.012)	0.975(.001)	**0.995(.000)**	**0.933(.024)**	0.959(.015)	0.967(.009)	**0.993(.002)**
	Real BCDR	0.912(.011)	0.946(.006)	0.968(.004)	0.994(.001)	0.885(.024)	0.931(.014)	0.935(.012)	0.987(.002)
Swin Transformer	None	0.834(.019)	0.891(.014)	0.928(.015)	0.986(.003)	0.784(.010)	0.860(.003)	0.877(.015)	0.976(.003)
	BCDR (WGAN-GP)	**0.950(.041)**	**0.969(.025)**	**0.978(.011)**	**0.996(.002)**	0.933(.020)	0.959(.013)	**0.973(.003)**	**0.995(.000)**
	BCDR (DCGAN)	0.876(.053)	0.920(.038)	0.959(.029)	0.992(.006)	0.922(.010)	0.953(.006)	0.968(.006)	0.994(.001)
	OPTIMAM (DCGAN)	0.941(.020)	0.964(.013)	0.975(.008)	0.995(.002)	0.920(.018)	0.952(.012)	0.963(.013)	0.993(.003)
	Both (a)	0.933(.018)	0.959(.012)	0.973(.012)	0.995(.002)	0.933(.010)	0.959(.007)	0.972(.002)	0.994(.000)
	Both (b)	0.914(.032)	0.947(.021)	0.972(.016)	0.995(.003)	**0.935(.010)**	**0.960(.006)**	0.970(.001)	0.994(.000)
	Real BCDR	0.841(.010)	0.907(.006)	0.929(.007)	0.987(.001)	0.799(.006)	0.862(.004)	0.890(.006)	0.979(.001)

Most importantly, all experiments with synthetic patches exhibit an improved performance of the classifier compared to using only INbreast. For example, when training a Swin Transformer on 50% of available INbreast training data and all lesion types, the classifier reaches an F1-score of 0.734. However, when adding synthetic patches from a DCGAN trained on BCDR and the same number of healthy samples from INbreast, the classifier reaches an F1-score of 0.880. Therefore, providing the classifier with synthetic data from another dataset improves its F1-score by 0.146 in this case. This effect tends to be strongest in the low-data regime, i.e. 50% of training data and only masses, and is particularly pronounced in the Swin Transformer.

4. DISCUSSION

In this work, we tested the hypothesis that sharing generative models instead of private patient data across centres is a beneficial alternative. We showed this empirically by comparing classification model performance when trained on synthetic data sampled from the shared generative models.

The results show that additional synthetic training data improves the performance of both classification models, where the improvement for the Swin Transformer is more pronounced than for the CNN. This is due to the larger number of trainable parameters in Swin Transformers (28M) compared to our CNN (1.1M). Additionally, since transformers do not possess inductive biases as CNNs, they require more training data to learn internal relations within the image. Therefore, the Swin Transformer is strongly benefiting from additional training data provided by the generators. In line with this argument the performance increase is most notable in the low-data regime, namely, where the training data is reduced by 50% and where only masses are used, as opposed to all lesion types.

Observing that adding the same number of synthetic patches from multiple sources, i.e. 50% BCDR and 50% OPTIMAM, does rarely improve upon adding 100% of a single-source (either BCDR or OPTIMAM). We hypothesize that as we increase the variation with multi-sources while leaving the dataset size constant, the classifier might have to learn/overcome additional domain shifts[28] and variation. For instance, OPTIMAM mass patches contain less background than BCDR mass patches while also being based on different acquistion protocols.

Furthermore, we note that the results often worsen when adding the real BCDR patches as compared to adding synthetic (BCDR) patches for both all lesions and only masses experiments. We suppose part of the performance decay stems from a domain-shift between BCDR and INbreast that could translate less into the GAN-generated BCDR synthetic data. We leave further investigation into this aspect to future work and point out that Garrucho et al[28] describe a considerable distribution-shift between BCDR and other mammography datasets, such as INbreast, in image contrast and intensity, as well as in the distribution of lesions, in terms of their size and aspect ratio.

As our results show promising potential of utilising GANs for privacy-preserving data sharing strategies, we also note that our proposed approach depends heavily on the willingness of centres to train and share generative models of their private datasets. In this context, we highlight the need for additional privacy preserving measures and thorough investigation as to how the shared generative model can leak private patient information (i) in general and (ii) when subjected to training data reconstruction attacks. One counter-measure against leakage of private information that we leave for future work is applying differential privacy[29] to GAN training, which provides a privacy guarantee for the GAN training data.[3] Before sharing a model, it is to be assessed whether private attributes such as particular patient-identifying anatomical or pathological features can be extracted from the shared model's parameters or predictions. For instance, model inversion attacks have shown to successfully reveal identifying imaging features from the training data.

To further validate the benefits of inter-centre generative model sharing, we recommend future work to test prediction performance improvement on full mammograms apart from patches and on further datasets across (imaging and non-imaging) modalities, organs, and standardised clinical tasks[6] with universal classification objectives (e.g., benign/malignant apart from healthy/non-healthy classification). We further propose future work to explore the effect of initialisation of classification models with pretrained weights alongside the effect of traditional preprocessing and data augmentation techniques, such as histogram and intensity scale normalization techniques.[28] Apart from classification, validation on further downstream tasks such as object detection, semantic segmentation, and domain-adaptation can reveal further insights into the advantages and limitations of generative model sharing. Lastly, further recommendations can be derived from the evaluation and comparison of additional types of GANs and generative models in diverse settings with varying patient data resource constraints between centres.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952103.

REFERENCES

[1] Global Cancer Observatory, "The global cancer observatory (gco) is an interactive web-based platform presenting global cancer statistics to inform cancer control and research." https://gco.iarc.fr/ (2021). Accessed: 2021-11-25.

[2] Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., and Abdel-Mottaleb, M., "Convolutional neural networks for breast cancer detection in mammography: A survey," *Computers in Biology and Medicine* **131**(January), 104248 (2021).

[3] Osuala, R., Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O., and Lekadir, K., "A review of generative adversarial networks in cancer imaging: New applications, new solutions," *arXiv preprint arXiv:2107.09543* (2021).

[4] Becker, A. S., Marcon, M., Ghafoor, S., Wurnig, M. C., Frauenfelder, T., and Boss, A., "Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer," *Investigative radiology* **52**(7), 434–440 (2017).

[5] Castro, D. C., Walker, I., and Glocker, B., "Causality matters in medical imaging," *Nature Communications* **11**(1), 1–10 (2020).

[6] Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., Aussó, S., Alberich, L. C., Marias, K., Tsiknakis, M., et al., "Future-ai: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging," *arXiv preprint arXiv:2109.09658* (2021).

[7] Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., et al., "Artificial intelligence in cancer imaging: clinical challenges and applications," *CA: a cancer journal for clinicians* **69**(2), 127–157 (2019).

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," in [*Advances in neural information processing systems*], 2672–2680 (2014).

[9] Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., and Michalski, M., "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in [*International workshop on simulation and synthesis in medical imaging*], 1–11, Springer (2018).

[10] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S., "Inbreast: toward a full-field digital mammographic database," *Academic radiology* **19**(2), 236–248 (2012).

[11] Lopez, M. G., Posada, N., Moura, D. C., Pollán, R. R., Valiente, J. M. F., Ortega, C. S., Solar, M., Diaz-Herrero, G., Ramos, I., Loureiro, J., et al., "Bcdr: a breast cancer digital repository," in [*15th International conference on experimental mechanics*], **1215** (2012).

[12] Halling-Brown, M. D., Warren, L. M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M. G., Wilkinson, L. S., Given-Wilson, R. M., McAvinchey, R., and Young, K. C., "Optimam mammography image database: A large-scale resource of mammography images and clinical data," *Radiology: Artificial Intelligence* , e200103 (2020).

[13] Fukushima, K., "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics* **36**(4), 193–202 (1980).

[14] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," *arXiv preprint arXiv:1706.03762* (2017).

[16] Radford, A., Metz, L., and Chintala, S., "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434* (2015).

[17] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A., "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028* (2017).

[18] Alyafi, B., Diaz, O., and Marti, R., "Dcgans for realistic breast mass augmentation in x-ray mammography," in [*Medical Imaging 2020: Computer-Aided Diagnosis*], **11314**, 1131420, International Society for Optics and Photonics (2020).

[19] Alyafi, B., Diaz, O., Elangovan, P., Vilanova, J. C., del Riego, J., and Marti, R., "Quality analysis of dcgan-generated mammography lesions," in [*15th International Workshop on Breast Imaging (IWBI2020)*], **11513**, 115130B, International Society for Optics and Photonics (2020).

[20] Osuala, R., Lazrak, N., Kushibar, K., Garucho, L., Jouide, S., Skorupko, G., Diaz, O., and Lekadir, K., "medigan: Synthetic Medical Data From Pretrained Generative Models," (Mar. 2022).

[21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030* (2021).

[22] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., "Pytorch: An imperative style, high-performance deep learning library," in [*Advances in Neural Information Processing Systems 32*], Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., eds., 8024–8035, Curran Associates, Inc. (2019).

[23] Magister, L. C. and Arandjelović, O., "Generative image inpainting for retinal images using generative adversarial networks," in [*2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*], 2835–2838, IEEE (2021).

[24] Arjovsky, M., Chintala, S., and Bottou, L., "Wasserstein generative adversarial networks," in [*International conference on machine learning*], 214–223, PMLR (2017).

[25] Agarap, A. F., "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375* (2018).

[26] Ioffe, S. and Szegedy, C., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in [*International conference on machine learning*], 448–456, PMLR (2015).

[27] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research* **15**, 1929–1958 (06 2014).

[28] Garrucho, L., Kushibar, K., Jouide, S., Diaz, O., Igual, L., and Lekadir, K., "Domain generalization in deep learning-based mass detection in mammography: A large-scale multi-center study," *arXiv preprint arXiv:2201.11620* (2022).

[29] Dwork, C., "Differential privacy," in [*Automata, Languages and Programming*], Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., eds., 1–12, Springer Berlin Heidelberg, Berlin, Heidelberg (2006).