Domain Adaptation for Medical Imaging

Saad Alrajhi, Drew Kulischak, Nick Russert, Francis Fernandez, Chao Jung Wu

Abstract

A core challenge of machine learning and artificial intelligence tasks is developing models that can be generalized to unseen data. Models that are trained and tested on the same source of data perform reasonably well. However, when models are trained on a particular domain or distribution, their performance tends to degrade when tested on other distributions. For medical imaging, the domain shifts can result from differences in scanning devices, scanning protocols, patient demographics, regions, and varying methods of data measurement and collection. One major obstacle in training medical image data is the scarcity of the data due to patient privacy, regulations, and costly data acquisition. One way to handle the limited availability of data is to add synthetic data to your training using generative models. Generative models are one popular way to address this issue because they have proven to be powerful at generating realistic-looking synthetic images. We hypothesize that adding synthetic data may make models more robust to out-of- domain distributions. We find that it is crucial to find models with vast enough differences in their methods to discover how these deep learning architectures perform with synthetic data added during training, and then compare them.

1. Introduction

1.1. Motivation

For machine learning tasks in medical imaging, the goal is often to determine whether the scan of the body part in the image has a specific condition or not. This is a fairly complex task because of how small details in medical images can make a difference on whether a patient appears to have a condition or not. Deep learning models have made great strides in the area of medical image classification, where models can discover neighboring and long-distance relationships in different parts of images and make predictions from there. Furthermore, it is likely that medical professionals will be more reliant on technology, and it is of great importance that this technology be can be depended on. Building sufficient models has real-world importance because of their potential to impact patients positively. Finding good datasets of medical images is challenging, as data is often scarce and constrained. Developing models that are able to generate real-looking synthetic medical images has two major benefits: improving models generalization for both within and across domains, and accelerating the data collection process.

1.2. Datasets

It is important to note the very small size of our datasets and how balanced they are for later observations in the paper. We define two distinct datasets that share the same modality, **greyscale ultrasound images of breast tumors**:

- **Egypt**: A collection of ultrasound images of breast tumors collected from Egyptian medical centers, 570 benign and 210 malignant. [1, 9].
- **Poland**: A collection of ultrasound images of breast cancer obtained from hospitals in Poland, 154 benign, 102 malignant [8]

The task is to classify whether the tumors in those images are benign (1) or malignant (0). For our training and synthetic data generation, we will be using the **Egypt** dataset, and then test on the **Poland** dataset. In other words, **Egypt** represents the in-domain data, while **Poland** represents the out-of-domain data.

1.3. Goals

Here are the aims of the project:

- Understand why certain models yield higher accuracies within the source domain.
- Build a model that is able to generate realistic looking fake images to add to our **Egypt** dataset that may increase the performance on out of distribution tests
- Discover what ratios of real and synthetic data from the **Egypt** domain yield the best results when tested on the **Poland** domain.

1.4. Contributions

For our achievements and experiments, we:

- modified Wasserstein Generative Adversarial Netowrk (WGAN) [2] model capable of creating realistic looking synthetic images from a small datasets
- built a CNN-Transformer inspired by TransMed that utilizes the strengths of both attention mechanisms and convolutions [3]
- Experiment and compare CNN-Transformer's perforamnce with Residual Networks [6] and Vision Transformers [4] on different ratios of real and synthetic data.

2. Related Work

2.1. Generative models

One common challenge in deep learning with medical images data, is the limited availability of said data. Privacy regulations and the costly acquisition of medical data has made it difficult for training models to learn from a limited data size. However, deep learning techniques, such as GANs, variational autoencoders, and diffusion models have it possible to get more training data [7]. The main goal of these models is to learn the underlying distribution of the data that is fed to them and generate images that closely match that distribution. For our project, we use GANs.

There are many variations of GAN models, so it is crucial to find a GAN model that generates medical images from a small dataset that look realistic and helps our training. We arrive at WGAN, a GAN model that uses a critic rather than a discriminator. GAN models often use a discriminator that judges fake images through a binary classification, however the WGANs use a score of the fake images instead. The WGAN uses a distance metric between the distributions of real and generated images called Earth mover's as the loss function, which offers great stability against Vanishing and exploding gradients:

$$L_{\text{critic}} = \max_{w \in \mathcal{W}} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))] \quad (1)$$

$$L_{\text{gen}} = \min_{\theta} \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))] \quad (2)$$

$$L_{\text{total}} = \min_{\theta} \max_{w \in \mathcal{W}} \mathbb{E}_{x \sim \mathbb{P}_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))] \quad (3)$$

The loss of the discriminator, also known as the critic, is the Earth Mover's distance and it is fed to the generator to train. The critic f_{w} learns to score images fed to it by the generator through evaluating how far the mean of the fake image is to the real image. This type of GAN works great in tasks that involve the generation of medical images, where data is scarce, variant, and a smoother gradient is crucially needed.

2.2. Domain Adaptation

The domain shifts in medical images are commonly caused by scanning protocols, use of different scanners and subject populations [5]. Models that are not able to adapt to differences between domains will have suboptimal results. Domain adaptation can be considered as a type of transfer learning, where the objective is to utilize both the feature space and the label space of a source domain and to get more accurate predictions on the target domains. Generating data from the source domain space while also considering its label and adding it to the training of a classification model may increase its performance on target domains. The assumption here is that synthetic data could bridge the gap between source and domain distributions by learning joint features through the augmented data.

3. Methods

3.1. Synthetic Data Generation

Finding an appropriate GAN model that specializes in generating realistic-looking synthetic images is difficult, especially for smaller datasets of ultrasound images. After experimenting with multiple GAN models, we found the Wasserstein GAN to be the most adequate in its performance.

Much of our WGAN code is a modified version of https://github.com/eriklindernoren/PyTorch-GAN/blob/master/implementations/wgan/wgan.py. We modify the code to add more convolutions in both the generator and discriminator, but also spectral-norm in the discriminator. The addition of more convolutions helps the generator and discriminator learn more about the smaller features of the image and helps classification models generalize better across domains. The spectral-norm stabilizes the learning of the discriminator from being too critical to smaller changes in input, and thus preventing the model from collapsing. We show some samples of both real and generated images and our progress in additional materials, (1) for real and (2,3,4,5,6) for synthetic images from the WGANs.

Since we are using a labeled dataset of benign and malignant images on our source domain of **Egypt**, we will generate both synthetic images based on real images labels and add the labels back to the generated

images. Then we add the synthetic data to the training set and perform the classification task on the target domain **Poland**. The pseudocode for this is in algorithm [1]:

Algorithm 1: Synthetic Data Generation

Input: $Eygpt_{real}$
Output: $Egypt_{mixed}$
Functions:
to generate synthetic data: G
to split data into benign and malignant: $Split$
concatenate sets: $Conc$
$Eygpt_{real-B}$, $Eygpt_{real-M} \leftarrow Split(\text{Eygpt}_{real})$
$Eygpt_{synth-B} \leftarrow G(\text{Eygpt}_{real-B})$
$Eygpt_{synth-M} \leftarrow G(\text{Eygpt}_{real-M})$
$Eygpt_{synth} \leftarrow Conc(\text{Eygpt}_{synth-M}, Eygpt_{synth-B})$
$Eygpt_{mixed} \leftarrow Conc(\text{Eygpt}_{synth}, Eygpt_{real})$

We preprocess the data into grayscale images of size 128x128, then we split the input of the whole dataset of **Egypt** into two: one set of benign and one set of malignant sets. After that, we generate synthetic images separately on the benign and the malignant images using the Wasserstein GAN. From there, we add both the synthetic datasets into a bigger synthetic dataset, where ratios of data could be taken. For our general task, we add all the synthetic data of **Egypt** with the real data of **Egypt** to get the mixed dataset. It is important to note that some of our synthetic data was created by simple horizontal flipping of the images, which has greatly helped in doubling the training data size of the mixed images.

3.2. Ratio tests

It is difficult to pinpoint the exact amount of synthetic data that may help or confuse the models, so we find it necessary to experiment with different ratios of real to synthetic data. Because of our constraints, we had to also trim down the dataset size for these tests to only a 100 for the real images of **Egypt** and at most a 100 for the synthetic images of **Egypt** in the training. In these experiments, we also use test results within the domain as a benchmark to measure how the accuracy of a trained model may be different when tested on both the source and target domains. In more detail, we train the models on real to synthetic ratios of: only real, 4:1, 2:1, 3:2, 1:1, and perform our tests on **Poland**. Furthermore, these experiments are implemented using our three models: Resnet, ViT, and a CNN-Transformer. We hope that these experiments helps us discover interesting characteristics of the models how they could learn or be confused by varying ratios of mixed data. Domain gaps are not the only measure to consider when testing which models may benefit from more synthetic data the most, we also need to consider how models generally perform on their source and target with the entirety of **Egypt** dataset with and without synthetic data.

3.3. Models

For our classification task, we make use of three models: ResNet50, ViT, and CNN-Transformer.

ResNets are very deep neural networks architectures that use residual connections. These connections are the basis of the residual blocks in the network, where inputs skip one or few layers and are added directly back to the output. We use a pre-trained model default Resnet-50 for as our convolutions-based model. Convolution-based models effectively learn neighbourhood features of the ultrasound images. Since we are training on a mix of synthetic and real data of **Egypt**, we guess that the Resnet's ability of learning smaller features of the image may benefit it when learning from synthetic images. If the WGAN generated images of **Egypt** look realistic enough for neighbourhood or smaller features, but also expand the learning of Resnet, then we suspect that adding the synthetic data to our training will lead to a higher accuracy when tested on the target domain **Poland**.

ViTs, which tokenize the input images into patches to find relationships between them, are great at discovering global and long-distance relationships between different parts of an image. Each of the patches represents a long embedding vectors of information that represents queries Q, Keys K, and Values, V.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (4)$$

Each query communicates its communication by its weight and a dot product with all the keys of the images, and from there each patch of the image is able to infer some information from all the other patches, which is the core of the transformers ability to capture relationships between all the patches. To combine both the strengths of the transform and convolutions models, we arrive at CNN-Transformer, which was inspired by the TransMed paper mentioned above. While TransMed was not specifically made for tasks involving classification of tumors of ultrasound images, its architecture may still offer valuable insights in performance within and across domains. We apply the ideas from the TransMed model by implementing a Convo-

	Tested on Source	Tested on Target	Domain Gap (400)
Trained on Egypt	80.62%	62.5%	17.76%
Trained on Poland	68.06%	49.01%	19.07%

Table 1. Accuracy and domain gap of all real data on ResNet model

lutional Neural Netowrk, which acts as the encoder of the model, then the encoded neighbourhood features are treated as inputs of the transformer. The transformer then treats each channel of feature outputs as a patch of 32x32, or 1024 patches in total to go through the multi-headed attention network. We hope through this method to combine both the strengths of convolutional networks that discover smaller neighbourhood features, but also utilize the transformers ability to learn long-distance relationships between a large amount of patches. Again, we hope that the CNN-Transformer model can learn both source domain distributions features from the real images of **Egypt** and the synthetic images would allow it to expand its learning to get a higher accuracy on the target domain of **Poland**

4. Evaluations

4.1. Results and Discussion

It is important to show how training and testing on different domains yields different results in terms of accuracy in the classification task. The test results in this section are average on about 3-5 tests.

In Tab. 1, we train and test on both domains and across the domains as well using the simple Resnet-50. Even though the images of both domains share the same modality of ultrasound, and the task is also the same, there is still a large gap performance across the domains. As previously mentioned, the difference between the domains or distributions in the data likely exists because of the different scans, scanning protocols, population, and regional differences. A crucial point to consider is that the **Poland** dataset is much smaller than the **Egypt** data as mentioned above, where it is only about a third of **Egypt**, but they share a similar ratio on their benign and malignant split. From this simple table, we conclude that getting more data is of great importance for target domain tests. **Egypt**'s target test is 13.49% higher than the **Poland** target domain test, likely just because **Poland**'s dataset is incredibly small. We also cannot say that domain gaps are the greatest measure of a models ability to be accurate across domains because that does not considers the models ability to get accurate results itself. An example of this would be: an accuracy of 50% on source domain and 49% across domains, here the model has only a 1% gap, but that does take into account the low accuracy of the model. For this reason, we consider general accuracy within and across domains to be the most important, but use the measure of domain gaps to understand how models behave differently on other domains.

As mentioned previously, we would like to discover how different real to synthetic images in our training may enhance our models' performance when tested on target domains. For this task, we use only a 100 images of our source domain **Egypt** and tested on its 100 real images from its domain as a bench mark, while also testing on different ratios of synthetic data and test on a 100 images of target domain **Poland**. We implement these tests of different ratios using a pretrained Resnet-50 in Tab. 2. Here the accuracy on the within domain test is 71.4% , but across the domain it drops to a 65%, and then we test the accuracy across the domain with different ratios of real to synthetic data. It seems that the Resnet-50 has this U-shaped results to its accuracy, where adding some synthetic data from **Egypt** using the WGAN confuses its training. However, when adding the most amount of synthetic data to the training, the result it a 66%. It could be that Convolution-based models learn from training data well, but do not benefit from a smaller amounts of synthetic data. In other words, it seems that convolution-based models learn out of distribution features when using larger amounts of synthetic data.

Now for our ViT tests in Tab. 3. We previously thought because transformer are the new-tech in deep learning tasks, they may be able to perform better than the convolution-based models, but that is not the case. The 16-patch ViT that we are using is falling behind the Resnet in the cross-domain tests greatly. The domain gap here is very relevant because we have about a 70% accuracy for test within the domain of **Egypt**, but the gap on the **Poland** tests is huge. The only point where we might see slight improvement is on 2:1 ratio of real to fake, where the accuracy is 58%. for the smallest amount of synthetic data 4:1, the synthetic data seem to be confusing the model, and in the other bottom two rows, the synthetic data makes no change.

for the ratio tests, the CNN-Transformer yields the best results on within domain test and perform much better than the transformer, and slightly better than

Train on	Real	Fake	Test on	Accuracy	Domain Gap
Real only	100	0	Egypt	71.4%	N/A
Real only	100	0	Poland	65%	6.4%
Real + Fake (4:1)	100	25	Poland	66%	5.4%
Real + Fake (2:1)	100	50	Poland	63%	8.4%
Real + Fake (3:2)	100	75	Poland	64%	7.4%
Real + Fake (1:1)	100	100	Poland	**66%**	5.4%

Table 2. Domain Gap and accuracy of ResNet50 model with various synthetic ratios

Train on	Real	Fake	Test on	Accuracy	Domain Gap
Real only	100	0	Egypt	70%	N/A
Real only	100	0	Poland	55%	15%
Real + Fake (4:1)	100	25	Poland	50%	20%
Real + Fake (2:1)	100	50	Poland	**58%**	12%
Real + Fake (3:2)	100	75	Poland	55%	15%
Real + Fake (1:1)	100	100	Poland	55%	15%

Table 3. Domain Gap and accuracy of Vision Transformer model with various synthetic ratios

Train on	Real	Fake	Test on	Accuracy	Domain Gap
Real only	100	0	Egypt	76.4%	N/A
Real only	100	0	Poland	63.6%	12.8%
Real + Fake (4:1)	100	25	Poland	65%	11.4%
Real + Fake (2:1)	100	50	Poland	65%	11.4%
Real + Fake (3:2)	100	75	Poland	**65.4%**	11%
Real + Fake (1:1)	100	100	Poland	65%	11.4%

Table 4. Domain Gap and accuracy of CNN-Transformer model with various synthetic ratios

Train Domain	Test Domain	ResNet50 Domain Gap	ViT Domain Gap	CNN-Transformer Domain Gap
Real only	Poland	6.4%	15%	12.8%
Real + Fake (4:1)	Poland	5.4%	20%	11.4%
Real + Fake (2:1)	Poland	8.4%	12%	11.4%
Real + Fake (3:2)	Poland	7.4%	15%	11%
Real + Fake (1:1)	Poland	5.4%	15%	11.4%

Table 5. Domain Gap of classification models with various synthetic ratios

Train set	Real (624)	Real (780)	Real (780) + Synthetic (400)	Real (780) + Synthetic (1580)
Test on	Egypt	Poland	Poland	Poland
Resnet	80.26%	61.3%	60.15%	**65.62%**
ViT	74.3%	60.5%	57.4%	60.2%
CNN-Transformer	**83%**	63.5%	62.27%	63.8%

Table 6. Accuracy of different datasets on classification models

the Resnet average Tab. 4. An interesting part of the CNN-Transformer, is that it benefits from synthetic data in its training at all stages. We compare the domain gaps of the three models in the Tab. 5. As mentioned, domain gaps are not our primary way to evaluate a model's performance across the domains, neither is using just ratios of data. For this reason, we have to consider how models perform in terms of just accuracy with all of **Egypt** in training as both real and a mixed set. Tab. 6. One of our goals was to build a model

that can classify tumors accurately within the source domain, which the CNN-Transformer does satisfyingly so at 83%. One interesting point is the three models seem to agree that when adding some synthetic data in the 4th column, the models seem to be getting confused in their leaning, but adding way more synthetic data, which some of it is flipped, improves the cross domain accuracy for the Resnet-50 and the CNN-Transformer.

It seems that models that are convolutions based benefit the most from synthetic data that is generated from smaller dataset in medical imaging classification tasks. Resnet here is learning the most from the synthetic data at 65.62%, even better that the fancy CNN-Transformer. Because the ViT transformer is not performing as well as models that utilize convolutions, the patching in the transformer may be learning attributes way too large in the medical images, and thus confusing the CNN-Transformer.

4.2. Limitations

As mentioned before, one of the most pressing challenges within the medical domain is data scarcity. Unlike many other computer vision datasets, medical image data are often limited due to the high cost for acquisition, the need for expert labeling, and sensitive patient information. GAN models are very data hungry and require a large number of training images in order to generate high quality synthetic images. This lack of data can also lead to models overfitting because there are so few examples for the model to learn general patterns from, even though we are using both dropout and a regularizer. Here is a sample of one of our training models (CNN-Transformer) from additional material (7). Since the training data can often be small for medical datasets, models may memorize the training dataset instead of learning the underlying structures, and thus perform sub-optimally.

Another common issue with medical data is that there is often a class imbalance. This can lead to important conditions being underrepresented within a dataset. An imbalance can bias a model towards the majority class and can hinder its ability to accurate predict on both labels, additional material (8). For real world Applications a biased model towards the benign classification would be highly problematic, because the recall on the malignant side would very high. If a tumor is malignant, you definitely don't want to classify it as benign.

Tasks of classifying medical image data, which often has some noise in them, is difficult. This is especially difficult when trying to get the data to get a better performance across domains. This problem has many layers to it that have to all be considered all the same, data scarcity, domain adaptation, model building, and high quality synthetic data generation.

4.3. Exploration and Future Work

Since convolution-based models seem to be performing better at this task than transformer based models, it might be worth it use more advanced convolutions models, such as Resnet-100,Resnet-152, Faster R-CNN, and others. It would likely be helpful for out project to utilize some sort of psuedo labeling methods based on some similarity metric between the images of the source and the target. It is also important to explore other domain adaptive methods with data augmentation such adversarial autoencoders and diffusion models, which as the literature suggests, could be promising.

5. Conclusion

We found that CNNs seemed to learn the most from the synthetic data and that it is very difficult to generate good synthetic images with such small datasets. To further that, its even harder to classify tumors as being malignant or benign within the same domain let alone another domain. Domain shifts are a very common problem in deep learning and in medical imaging tasks as well. There is great demand for real-world applications in health related tasks, and machine learning models that make accurate predictions within and across domains are desperately needed. To address data scarcity in medical imaging, we utilize generative models that specialize in our task and have achieved satisfying results on our tasks. This solves both the problem of data scarcity, but also expedites the process of gather personal and sensitive patient data while also showing real potential in helping machine learning models generalize to unseen distributions. Testing our models with different ratios enables to understand how accuracies may change across domains. It is also important to test our models with the full dataset. It is worth it to explore more data augmentation techniques and modifications from the literature and consider newer methods of domain adaptation methods, not just in medical imaging.

Contributions:
Saad Alrajhi: BIG-WGAN, CNN-Transformer, Final paper
Drew Kulischak: ResNet, Vision Transformer
Nick Russert: Gan Exploration, Synthetic Data
Francis Fernandez: ResNet, GANs
Chao Jung Wu: ResNet, ViT

6. References

[1] Wafaa Al-Dhabyani, Maha Gomaa, Reda Khaled, and Abeer Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 2020. 1

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *Proceedings of Machine Learning Research (PMLR)*, 2017. 2

[3] Ying Dai and Yang Gao. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 2021. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2020. 2

[5] Haoran Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 2022. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2

[7] Ahmed Kebaili, Thouraya Slimani, and Neila Mbarek. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 2023. 2

[8] Agnieszka Pawłowska, Marcin Topolski, Bartosz Świderski, et al. Curated benchmark dataset for ultrasound-based breast lesion analysis. *Scientific Data*, 2024. 1

[9] Jieneng Yang, Jiancheng Jiao, Huiyu Zhang, et al. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 2023. 1