Karaelmas Science and Engineering Journal
Journal home page: https://dergipark.org.tr/tr/pub/karaelmasfen
DOI: 10.7212/karaelmasfen.1489168

Research Article
Received / Geliş tarihi : 27.05.2024
Accepted / Kabul tarihi : 15.09.2024

Generation of Synthetic Data Using Breast Cancer Dataset and Classification with Resnet18

Meme Kanseri Veri Seti Kullanılarak Sentetik Veri Üretilmesi ve Resnet18 ile Sınıflandırılması

Dilşat Berin Aytar*, Semra Gündüç

Ankara University, Department of Computer Engineering, Ankara, Türkiye

Abstract

Since technology is advancing so quickly in the modern era of information, data is becoming an essential resource in many fields. Correct data collection, organization, and analysis make it a potent tool for successful decision-making, process improvement, and success across a wide range of sectors. Synthetic data is required for a number of reasons, including the constraints of real data, the expense of collecting labeled data, and privacy and security problems in specific situations and domains. For a variety of reasons, including security, ethics, legal restrictions, sensitivity and privacy issues, and ethics, synthetic data is a valuable tool, particularly in the health sector. A Deep Learning (DL) model called GAN (Generative Adversarial Networks) has been developed with the intention of generating synthetic data. In this study, the Breast Histopathology dataset was used to generate malignant and benign labeled synthetic patch images using MSG-GAN (Multi-Scale Gradients for Generative Adversarial Networks), a form of GAN, to aid in cancer identification. After that, real and synthetic data were classified in four different ways with Transfer Learning (TL) using the ResNet18 model. In the first classification, real data is used as training and test data and an accuracy rate of 84%, in the second classification, synthetic data is used as training and test data and the accuracy rate is 99%, in the third classification, real data is used as training and synthetic data is used as test data and an accuracy rate of 81%, in the fourth classification, synthetic data is used as training and real data is used as test data and an accuracy rate of 76%. As a result of the study, four different classifications were associated and it was tried to determine whether the synthetic images are similar to the original data and whether they behave like real data.

Keywords: Generative adversarial networks, histopathology, MSG-GAN, ResNet18, synthetic data.

Öz

Bilgi çağı olan günümüzde veri, özellikle teknolojinin hızla ilerlemesiyle birçok alanda kritik bir kaynak hâline gelmiştir. Veri doğru bir şekilde toplandığında, düzenlendiğinde ve analiz edildiğinde birçok sektörde etkili kararlar almak, süreçleri iyileştirmek ve başarı elde etmek için güçlü bir araç hâline gelir. Gerçek verinin kısıtlılığı, etiketlenmiş verinin elde edilmesinin maliyetli olması, bazı durumlarda ve alanlarda gizlilik ve güvenlik endişeleri gibi sebepler sentetik verilere ihtiyaç duyulmasına sebep olmuştur. Sentetik veriler, özellikle sağlık alanında hassaslık ve gizlilik endişeleri, yasal düzenlemeler, etik ve güvenliğin sağlanmaya çalışılması gibi nedenlerden dolayı önemli bir araçtır. Sentetik veri üretme amacıyla Derin Öğrenme (DÖ) modeli olan ÇÜA (Çekişmeli Üretici Ağlar) ortaya çıkmıştır. Bu çalışmada Meme Histopatoloji veri seti kullanılarak bir ÇÜA çeşidi olan ÇÖD-ÇÜA (Üretken Rekabetçi Ağlar için Çok Ölçekli Değişimler) ile kanser tespitinde yarar sağlamak amacıyla kötü huylu ve iyi huylu etiketli sentetik yama görselleri oluşturulmuştur. Sonrasında gerçek ve sentetik veriler ResNet18 modeli kullanılarak Aktarımlı Öğrenme ile dört farklı şekilde sınıflandırılmıştır. İlk sınıflandırmada gerçek veriler eğitim ve test verisi olarak kullanılıp %84 doğruluk oranı, ikinci sınıflandırmada sentetik veriler eğitim ve test verisi olarak kullanılıp %99 doğruluk oranı, üçüncü sınıflandırmada gerçek veriler eğitim, sentetik veriler test verisi olarak kullanılıp %81 doğruluk oranı, dördüncü sınıflandırmada sentetik veriler eğitim, gerçek veriler test verisi olarak kullanılıp %76 doğruluk oranı elde edilmiştir. Çalışma sonucunda dört farklı sınıflandırma ilişkilendirilerek sentetik görüntülerin orijinal verilere olan benzerliği ve gerçek veri gibi davranıp davranmadığı tespit edilmeye çalışılmıştır.

Anahtar Kelimeler: Çekişmeli üretici ağlar, histopatoloji, ÇÖD-ÇÜA, ResNet18, sentetik veri.

*Corresponding author: aytar.berin@gmail.com

Dilşat Berin Aytar orcid.org/0009-0006-9984-241X
Semra Gündüç orcid.org/0000-0002-3811-9547

1. Introduction

In today's world, data has become an extremely important resource in many fields, akin to a valuable commodity. The usability of data in numerous areas has led to the emergence of new data and an increased need for more data. Fields such as scientific research, decision-making processes, strategic planning, performance monitoring, Artificial Intelligence (AI) and Machine Learning (ML), customer experience and personalization, scientific research and innovation, healthcare services and medicine, security and risk management, solving societal issues, and marketing demonstrate the significance of data.

Data plays many important roles in the field of healthcare and medicine, contributing significantly to the development of healthcare services, patient treatment, and the formation of health policies. However, there are limitations to accessing real data in healthcare. Health data often contains sensitive and confidential information. The sharing and access of medical data are subject to strict regulations and can be limited due to privacy concerns, ethics, and security. Examples of legal regulations governing health data include HIPAA (Health Insurance Portability and Accountability Act) in the United States, GDPR (General Data Protection Regulation) in the European Union, and various other regulations in different countries related to the protection and processing of health data. Furthermore, health data is protected under medical ethical rules and standards. These principles address issues such as data confidentiality, patient privacy, and patient rights, guiding healthcare providers. Therefore, healthcare institutions and providers must strictly adhere to these regulations and take various measures to ensure the confidentiality and security of health data, ensuring patient safety and preventing data misuse. Overcoming limitations in accessing data is crucial for health research and innovations.

Overcoming limitations and addressing data scarcity, synthetic data generated by AI presents a significant solution. Synthetic data, created artificially by a computer program, is designed to mimic the characteristics of real-world data while preserving individual privacy and avoiding data breaches. Organizations can generate nearly unlimited amounts of data for testing, research, and analysis using synthetic data without worrying about ethical and legal issues associated with real-world data. Synthetic data, generated through advanced algorithms and models, offers a viable solution to these challenges by creating artificial datasets that mimic the statistical properties of real-world data. This enables researchers to augment existing datasets, perform robust experiments, and train ML models more effectively (Goodfellow et al. 2014).

Generative Adversarial Networks (GANs) have emerged for the purpose of synthetic data generation. GANs, introduced by Goodfellow et al. in 2014, consist of two neural networks-a generator and a discriminator-that compete against each other to produce realistic data samples. This adversarial training process results in highly realistic synthetic data that can be used in various applications, including image classification, natural language processing, and anomaly detection (Goodfellow et al. 2014). Recent advancements in GANs have led to the development of more sophisticated models, such as StyleGAN and CycleGAN, which further enhance the quality and diversity of synthetic data (Karras et al. 2019, Zhu et al. 2017).

GANs are a significant and innovative modeling approach in the field of DL. They consist of two networks, a generator and a discriminator, which compete with each other. These networks compete to ultimately produce realistic images. For example, GAN-generated medical images have been used to improve the accuracy of diagnostic models and to train models on rare diseases where real data is scarce (Frid-Adar et al. 2018, Salehinejad et al. 2018). Various types of GAN model variants have been developed to meet different needs. This article will describe Multi-Scale Gradients for Generative Adversarial Networks (MSG-GAN), a type of GAN model.

In MSG-GAN, the generator and discriminator networks compete at a single resolution and improve together. The MSG technique utilizes different resolution levels to stabilize this competition. This approach gradually increases operations starting from lower resolutions and scales up to real dimensions. This innovation addresses common issues in GAN training, such as mode collapse and training instability, leading to more robust synthetic data generation (Karnewar et al. 2019). As a result, a faster, more stable, and improved training process is provided, contributing to more realistic, consistent, and high-quality results by better utilizing information at different scales. MSG-GAN introduces a multi-scale gradient approach that enables the generator to produce high-resolution images with finer details by receiving gradients at multiple scales during the training process. MSG-GAN is particularly successful in tasks such as image generation and synthesis involving visual datasets. It has a wide range of applications in data augmentation, AI studies, art, computing, medicine, automotive, finance, and many other fields.

After synthetic data generation with MSG-GAN, the generated synthetic data will be classified using TL techniques such as ResNet18 (Residual Neural Network). ResNet18 is a DL model commonly used for visual recognition and classification problems. The term "ResNet" stands for "Residual Networks," and "18" denotes the number of layers in the model.

ResNet represents an architecture that includes residual blocks developed to facilitate the training of deep neural networks and reduce overfitting. The residual block passes the input data through an activation function and several convolutional and summation layers. An important feature of ResNet is the presence of "residual connections" in these blocks.

TL refers to the reuse of features learned by a pre-trained model to solve a different task. If necessary, weights are adjusted and new layers are added based on the new task and dataset. For example, a pre-trained ResNet18 model may have been trained for many image classification tasks. In a new image classification task, the pre-trained ResNet18 model can be taken and retrained on the new dataset.

Breast cancer is the second most common cancer type globally after lung cancer (Teh et al. 2015). Invasive Ductal Carcinoma (IDC) is the most common subtype among all breast cancers. The aim is to reduce reliance on pathologists and thereby reduce errors and human-related biases during disease detection, as well as minimize the high economic cost and time loss associated with it. In this study, IDC+ and IDC- histopathological images will be generated using MSG-GAN for disease detection, and the images will be classified using ResNet18.

2. Material and Methods

In the two-stage study, in the first stage, synthetic images were produced using MSG-GAN, and in the second stage, classification was made using ResNet18, one of the TL techniques. Finally, the classification results were evaluated with metrics.

2.1. MSG-GAN (Multi-Scale Gradients Generative Adversarial Network)

MSG-GAN (Karnewar et al. 2019), is a technique used to enhance the performance of traditional GANs (Goodfellow et al. 2014) by stabilizing their training process and achieving high-quality results.

The MSG technique uses different resolution levels (starting from lower to higher resolutions) to stabilize this competition. This approach begins with lower resolutions and gradually scales up operations to the actual size. This ensures a faster, more stable, and improved training process, leveraging information at different scales to produce more realistic, consistent, and high-quality results.

The process generally involves the following steps:

- **Start at Low Resolution:** Initially, the generator network operates at a lower resolution, learning simpler patterns.
- **Increase Resolution:** As the generated images are scaled to higher resolutions, the complexity of the network increases, adding more detail and realism.
- **Finalize at Real Size:** The process continues until the target resolution is reached, allowing the network to learn more complex patterns and details.

This technique is considered a significant advancement in the evolution of GANs, often leading to better performance on high-resolution images or other complex data types. MSG-GAN is utilized in various image synthesis and other application areas.

Figure 1 shows the basic MSG-GAN architecture used in the study. The architecture includes connections from the intermediate layers of the generator to the intermediate layers of the discriminator. The multi-scale images sent to the discriminator are combined with activation volumes obtained from the main path of the convolutional layers using a "Combine Function" (shown in yellow) (Karnewar et al. 2019).

2.1.1. MSG-GAN Generator Architecture and Function

The generator architecture used to produce a 64x64x3 image with MSG-GAN typically consists of 5 blocks. While Table 1, illustrates the entire generator architecture, the 5 blocks used to generate 64x64x3 images in the study are highlighted in bold.

The generator architecture generally involves upsampling and convolutional operations. Upsampling refers to transforming images into higher resolutions. After each upsampling step, two convolution operations are usually performed. The generator processes input noise to generate realistic images with dimensions of 64x64x3.

The generator typically takes random noise vectors as input. This noise is processed and transformed into a feature map containing pixel values. After upsampling the images, the

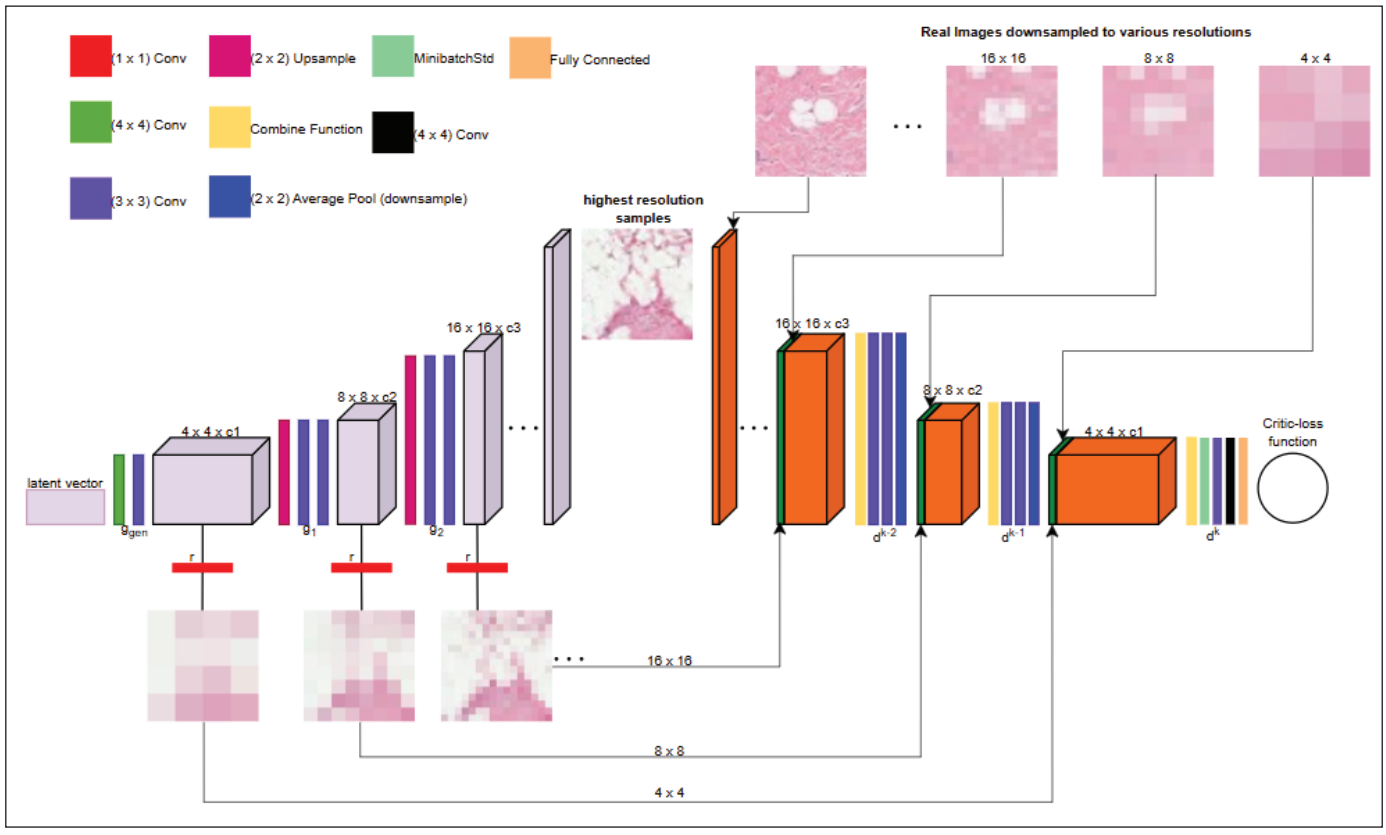

Figure 1. MSG-GAN architecture.

generator typically learns features using convolutional layers with 3x3 filters. Higher-level features are learned at each layer. Initially, these features may represent simple patterns and shapes, which later evolve into more complex objects and structures. The LeakyReLU activation function (Xu et al. 2015) is used in each block.

The generator performs two transmission operations in each block:

- Firstly, after every second convolution operation, the generator passes its output to the next block until reaching the final block. The output size, for example, is 512x16x16.
- Secondly, in addition to Table 1, a 1x1 convolution operation is applied to the output of the respective block before transitioning to the next block. This ensures that the block output has 3 channels (RGB features). This output is transmitted as input to the discriminator. The output size, for example, is 3x16x16.

By following the specified steps and performing the 1x1 convolution operation, the generator produces high-quality images of size 64x64x3 in five blocks and transmits them to the discriminator.

The generator is trained using feedback from the discriminator. The discriminator evaluates the realism of the generated images and provides feedback. The generator learns to make the generated images increasingly realistic based on this feedback. During the training process, the generated images are optimized to resemble real images as closely as possible. This helps prevent issues such as mode collapse and training instability

2.1.2. MSG-GAN Discriminator Architecture and Function

The discriminator architecture used to generate a 64x64x3 image with MSG-GAN typically consists of 5 blocks. Table 2 displays the entire discriminator architecture, with the 5 blocks highlighted in bold for generating the 64x64x3 images in the study. The 5 blocks used in the discriminator architecture are the last 5 blocks compared to the first 5 blocks used in the generator architecture. This is because, as described below, the output of each block from the generator will be the input to the corresponding blocks of the discriminator from end to start.

The discriminator model is structured to handle images of different sizes in each block. Each block in Table 2 represents a different scale level. Block operations typically involve taking the raw RGB image, concatenating it with feature maps from previous blocks (Concat/φ_simple), adding minibatch standard deviation (MiniBatchStd) to the feature maps, applying a 3x3 convolution operation (a 3x4 convolution operation is applied in the last block), and performing average pooling (Avg Pooling). Each convolution layer uses a certain number of filters, and the LeakyReLU activation function (Xu et al. 2015) is used.

Table 1. Generator architecture.

Block	Operation	Activation Function	Output Shape
1.	**Latent vector** **Conv 4x4** **Conv 3x3**	**Norm** **LReLU** **LReLU**	**512x1x1** **512x4x4** **512x4x4**
2.	**Upsample** **Conv 3x3** **Conv 3x3**	**-** **LReLU** **LReLU**	**512x8x8** **512x8x8** **512x8x8**
3.	**Upsample** **Conv 3x3** **Conv 3x3**	**-** **LReLU** **LReLU**	**512x16x16** **512x16x16** **512x16x16**
4.	**Upsample** **Conv 3x3** **Conv 3x3**	**-** **LReLU** **LReLU**	**512x32x32** **512x32x32** **512x32x32**
Model 1 ↑			
5.	**Upsample** **Conv 3x3** **Conv 3x3**	**-** **LReLU** **LReLU**	**512x64x64** **256x64x64** **256x64x64**
6.	Upsample Conv 3x3 Conv 3x3	- LReLU LReLU	256x128x128 128x128x128 128x128x128
Model 2 ↑			
7.	Upsample Conv 3x3 Conv 3x3	- LReLU LReLU	128x256x256 64x256x256 64x256x256
Model 3 ↑			
8.	Upsample Conv 3x3 Conv 3x3	- LReLU LReLU	64x512x512 32x512x512 32x512x512
9.	Upsample Conv 3x3 Conv 3x3	- LReLU LReLU	32x1024x1024 16x1024x1024 16x1024x1024
Model full ↑			

In the last block, average pooling is not performed. Instead, there is a fully connected layer. The fully connected layer produces an output to determine whether an input image of size 64x64x3 is real or fake. During training, it attempts to produce high output values for real images and low output values for generated images as much as possible.

2.2. ResNet18

ResNet (He et al. 2015) is a neural network model introduced to facilitate the training of deep networks and

Table 2. Discriminator architecture.

Block	Operation	Activation Function	Output Shape
Model full ↓			
1.	Raw RGB images 0 FromRGB 0 MiniBatchStd Conv 3x3 Conv 3x3 AvgPool	- - - LReLU LReLU -	3x1024x1024 16x1024x1024 17x1024x1024 16x1024x1024 32x1024x1024 32x512x512
2.	Raw RGB images 1 Concat/ MiniBatchStd Conv 3x3 Conv 3x3 AvgPool	- - - LReLU LReLU -	3x512x512 35x512x512 36x512x512 32x512x512 64x512x512 64x256x256
Model 3 ↓			
3.	Raw RGB images 2 Concat/ MiniBatchStd Conv 3x3 Conv 3x3 AvgPool	- - - LReLU LReLU -	3x256x256 67x256x256 68x256x256 64x256x256 128x256x256 128x128x128
Model 2 ↓			
4.	Raw RGB images 3 Concat/ MiniBatchStd Conv 3x3 Conv 3x3 AvgPool	- - - LReLU LReLU -	3x128x128 131x128x128 132x128x128 128x128x128 256x128x128 256x64x64
5.	**Raw RGB images 4** **Concat/** **MiniBatchStd** **Conv 3x3** **Conv 3x3** **AvgPool**	**-** **-** **-** **LReLU** **LReLU** **-**	**3x64x64** **259x64x64** **260x64x64** **256x64x64** **512x64x64** **512x32x32**

Table 2. Cont.

Block	Operation	Activation Function	Output Shape
Model 1 ↓			
6.	**Raw RGB images 5**	**-**	**3x32x32**
	Concat/	**-**	**515x32x32**
	MiniBatchStd	**-**	**516x32x32**
	Conv 3x3	**LReLU**	**512x32x32**
	Conv 3x3	**LReLU**	**512x32x32**
	AvgPool	**-**	**512x16x16**
7.	**Raw RGB images 6**	**-**	**3x16x16**
	Concat/	**-**	**515x16x16**
	MiniBatchStd	**-**	**516x16x16**
	Conv 3x3	**LReLU**	**512x16x16**
	Conv 3x3	**LReLU**	**512x16x16**
	AvgPool	**-**	**512x8x8**
8.	**Raw RGB images 7**	**-**	**3x8x8**
	Concat/	**-**	**515x8x8**
	MiniBatchStd	**-**	**516x8x8**
	Conv 3x3	**LReLU**	**512x8x8**
	Conv 3x3	**LReLU**	**512x8x8**
	AvgPool	**-**	**512x4x4**
9.	**Raw RGB images 8**	**-**	**3x4x4**
	Concat/	**-**	**515x4x4**
	MiniBatchStd	**-**	**516x4x4**
	Conv 3x3	**LReLU**	**512x4x4**
	Conv 3x4	**LReLU**	**512x1x1**
	Fully Connected	**Linear**	**1x1x1**

increase performance in the field of visual processing. The main purpose of ResNet is to reduce training difficulties that may occur by making the network deeper.

ResNet uses the concept of residual learning. In this approach, the network tries to learn the difference between the input data and the output, instead of just predicting the outputs of the layers. The basic structural unit of ResNet is the Residual Block. This block is slightly different from a traditional neural network layer. A Residual Block consists of convolution, activation, normalization layers and Residual Connection (Shortcut). Residual connections, which ease the training of deep neural networks, enhance their performance, and enable deeper networks.

ResNet is named according to the number of layers used. In this study, the ResNet18 model with 18 layers was used. ResNet18, a pre-trained TL model, is used to perform binary classification in this study.

ResNet18 comprises convolutional layers, batch normalization, and ReLU activation functions (Liu 2020), Residual connections involve adding the output of the previous layer to the next layer. This technique mitigates the vanishing gradient problem and allows the training of very deep networks possible.

ResNet18 and other ResNet architectures are applied in many areas such as image processing, object recognition, face recognition and medical image analysis. In particular, TL techniques and pre-trained networks such as ResNet18 are used to achieve high success rates in various tasks.

The reasons for choosing the ResNet18 model in this project are based on various factors. First, the large dataset size supports the use of a medium-scale model such as ResNet18. While this model can provide sufficient performance on a data set of 160,000 samples, it can optimize training times and memory usage by having a lighter structure than deeper networks. Moreover, thanks to its ability to capture general data patterns, it can be used effectively in various scenarios of the project (e.g. combinations of real and synthetic data). As a result, the ResNet18 model was evaluated as a suitable choice to achieve the balance of scale and performance required by working with large data sets.

In the study, all layers of the pre-trained model except the last layer are frozen and will not be updated during training. Only the last Full Connection layer of the model was changed and the output was arranged to be two classes (positive or negative). The parameters (weights and bias) of the last added Full Connectivity layer are set to be open to training. Figure 2 shows the ResNet18 architecture used in the study.

2.3. Dataset and Summary of Work

Firstly, the training dataset was downloaded from https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images. Janowczyk et al. 2016, Cruz-Roa et al. 2014). Patches of IDC, the most prevalent subtype of breast cancer, are seen in the dataset. Of all breast cancers, invasive ductal carcinoma (IDC) is the most prevalent subtype. Pathologists normally concentrate on areas with IDC when determining the aggressiveness levels of each assembly sample. Thus, locating the precise IDC zones across the entire assembly slide is a typical preprocessing step for automatic aggressiveness rating. There are 156,000 patches in all, consisting of 78,000 IDC negative and 78,000 IDC positive patches. The pictures have three channels and a 50x50 dimension. Data that is IDC negative is labeled 0, whereas data that is IDC positive is labeled 1.

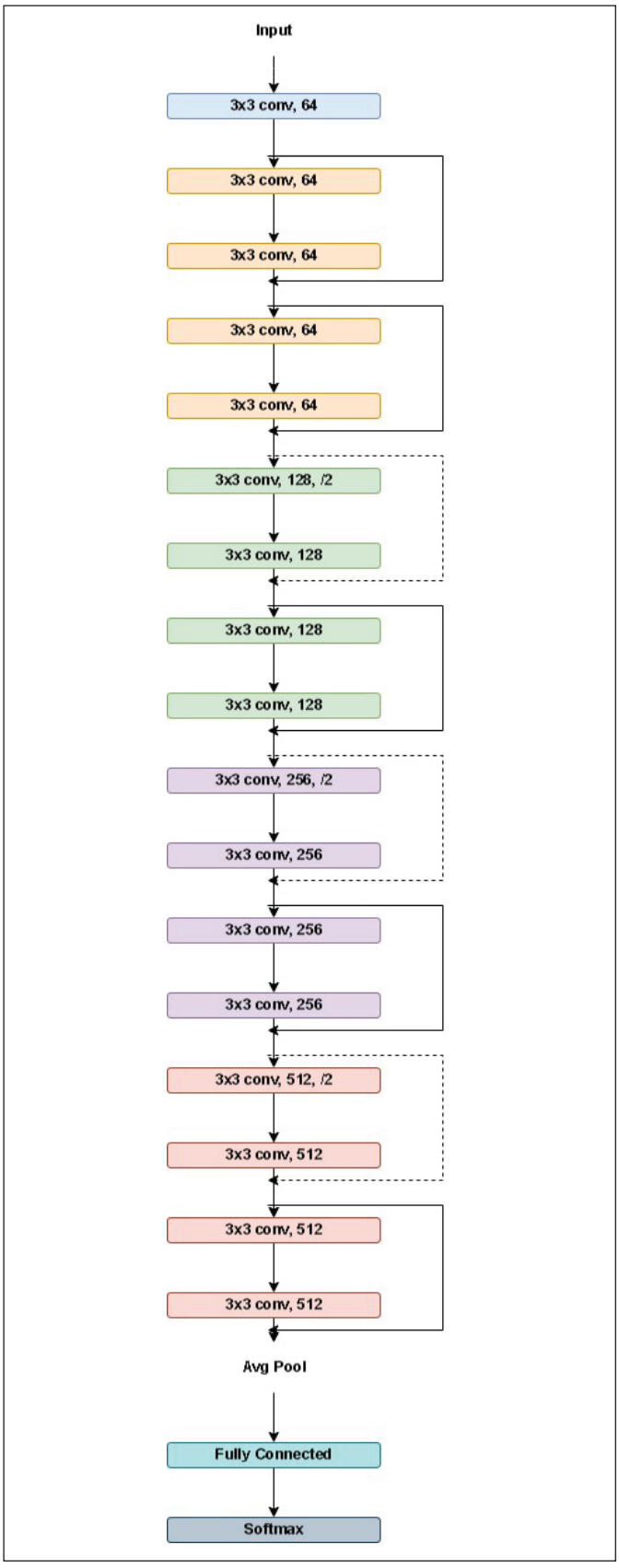

Figure 2. ResNet18 architecture.

The dataset was divided into two parts for synthetic data generation (40,000 x 2) and classification (38,000 x 2). The reason for this is to ensure that the test data consists of previously unseen data during the training phase when performing classification.

The study consists of two parts.

- Firstly, synthetic data generation was performed using MSG-GAN with 80,000 images. As a result of this study, 38,000 synthetic IDC negative and 38,000 synthetic IDC positive data were generated.
- Secondly, four different classifications were performed using ResNet18, which is one of the TL models.

2.4. Environmental Variables Used

All models used in this study were compiled with GPU/CPU support. All codes are implemented with the PyTorch 2.0 framework, an open-source deep neural network library written in Python.

3. Results and Discussion

A certain number of datasets were divided into two parts, one used to generate breast cancer negative and positive labeled images using MSG-GAN. The generated images and the unused portion of the real data were used as training and test data, and four different classification processes were performed using ResNet18. The model was evaluated based on the results obtained.

3.1. Synthetic Data Generation with MSG-GAN

As explained in section 2.1.1 and section 2.1.2, a 5-block generator and discriminator architecture was used with MSG-GAN. In the study, WGAN-GP (Gulrajani et al. 2017) was determined as the loss function for both the generator and discriminator models. As a result, 38,000 synthetic IDC- and 38,000 synthetic IDC+ data of size 64x64x3 were produced by using 40,000 IDC+ and 40,000 IDC- data from the real data set. The hyperparameters used in each of the classifications are as follows: Batch size is 16, learning rate is 0.0001, and optimizer is RMSprop (Graves 2013).

Figure 3 shows examples of synthetic IDC+ data produced using MSG-GAN, and Figure 4 shows examples of synthetic IDC- data produced using MSG-GAN.

38,000 synthetic IDC+ and 38,000 synthetic IDC- data produced using MSG-GAN were used in the next stage of the study, namely classification.

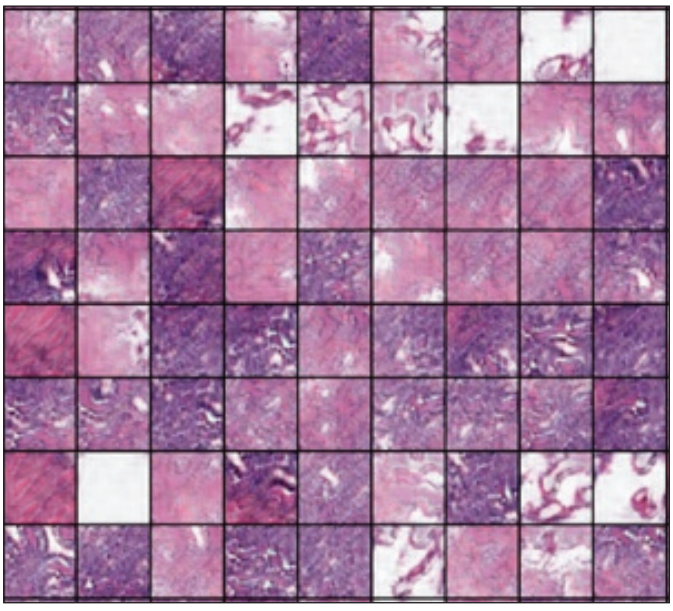

Figure 3. Synthetic IDC+ data examples.

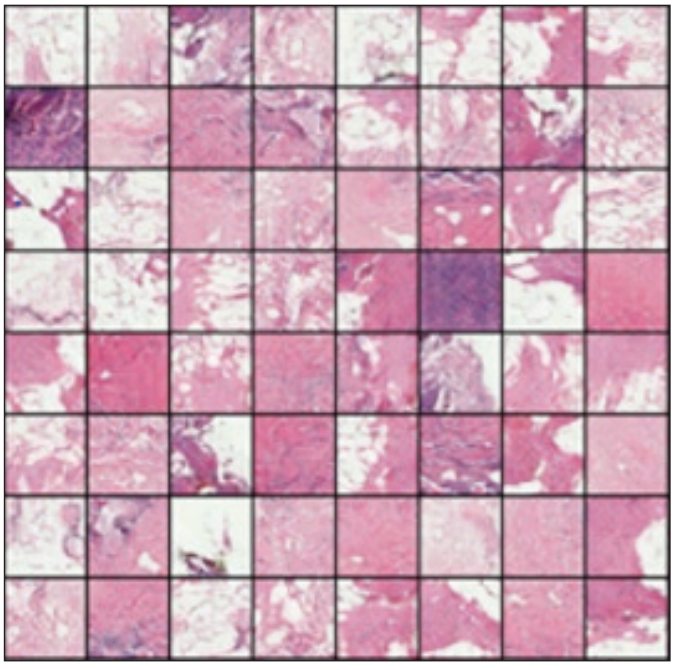

Figure 4. Synthetic IDC- data examples.

3.2. Classification of Real and Synthetic Data with ResNet18

In this stage, classification was performed using the unused 76,000 portion of the real dataset and the generated 76,000 synthetic data with the pre-trained ResNet18 model.

The purpose of this section was to determine how similar synthetic data was generated to the real data and to classify images. This section was divided into four sub-sections. The same model was used in each subsection. The subsections are as follows:

- Training and classification of real data: 70% (53,200 patches) of the real data was used as training and 30% (22,800 patches) as testing. (Even though the data set was divided into training and test data in the ratios of 80:20, 60:40, 50:50, the result did not change. For this reason, the division is shown as 70:30.)
- Training and classification of synthetic data: 70% (53,200 patches) of the synthetic data was used as training and 30% (22,800 patches) as testing. (Even though the data set was divided into training and test data in the ratios of 80:20, 60:40, 50:50, the result did not change. For this reason, the division is shown as 70:30.)
- Training with real data and classification of synthetic data: All real data (76,000 patches) was used as training and all synthetic data (76,000 patches) was used as test data.
- Training with synthetic data and classification of real data: All synthetic data (76,000 patches) was used as training data and all real data (76,000 patches) was used as test data.

The purpose of the four different classifications is to establish a relationship between the classification results, examine the ability of synthetic data to reflect real data, and determine whether synthetic data behaves like real data.

A DL model like ResNet18 generally expects an input size of 224x224x3. Therefore, in the study, real and synthetic data were resized to 224x224x3 dimensions before the classification process. The ResNet18 model architecture described in Section 2.2 was implemented.

CrossEntropyLoss (LeCun et al. 1998) was used as the loss function. The hyperparameters used in each of the classifications are as follows: Batch size is 32, learning rate is 0.001, and optimizer is Adam (Kingma and Ba, 2015).

Training lasted for 150 epochs for each classification. Loss was calculated for each batch, gradients were propagated backward, and model parameters were updated using the optimizer.

3.3. Evaluation of The Classification Results

To evaluate the system's performance, accuracy, precision, recall, and F1 score metrics were used. Table 3 shows the metrics obtained from four different classifications conducted in the study.

Table 3. Metrics obtained as a result of classification of real and synthetic data.

Train/Test Data	Accuracy	Precision	Recall	F1 Score
Real/Real	0.84	0.84	0.84	0.84
Synthetic /Synthetic	0.99	0.98	0.98	0.98
Real / Synthetic	0.81	0.82	0.78	0.78
Synthetic / Real	0.76	0.77	0.76	0.76

Firstly, achieving high accuracy, precision, recall, and F1 score values when using real data as both training and test data (Real/Real) for classification indicates that the pre-trained ResNet18 model can be considered as a baseline model for this study. These results reflect the model's ability to classify real-world data accurately. Figure 5 shows the ROC curve of the classification and Table 4 shows the confusion matrix of the classification.

Table 4. Confusion matrix (Percentage).

		Actual	
		IDC+	IDC-
Predicted	**IDC+**	84.1	15.9
	IDC-	15.8	84.2

The lower accuracy, precision, recall, and F1 score values obtained when using real data for training and synthetic data for testing (Real/Synthetic) compared to Real/Real results can be attributed to the fact that the distribution conformity in synthetic data does not perfectly match that of real data, leading to out of distribution data samples in the generated data. These data samples decrease the accuracy rate of the respective classification. However, the similarity between the metric values obtained from Real/Real and Real/Synthetic classifications indicates that synthetic data closely resemble real data. Figure 6 shows the ROC curve of the classification and Table 5 shows the confusion matrix of the classification.

Table 5. Confusion matrix (Percentage).

		Actual	
		IDC+	IDC-
Predicted	**IDC+**	82.5	17.5
	IDC-	21.2	78.8

Achieving nearly 100% accuracy, precision, recall, and F1 score values when using synthetic data as both training and test data (Synthetic/Synthetic) suggests that the model's capacity is sufficient to learn the distribution of the data. These high rates indicate that synthetic data can be easily learned by the model. Figure 7 shows the ROC curve of the classification and Table 6 shows the confusion matrix of the classification.

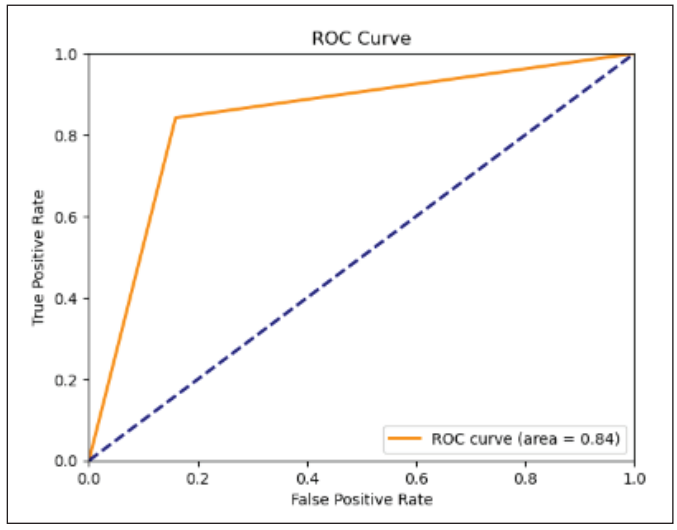

Figure 5. ROC curve.

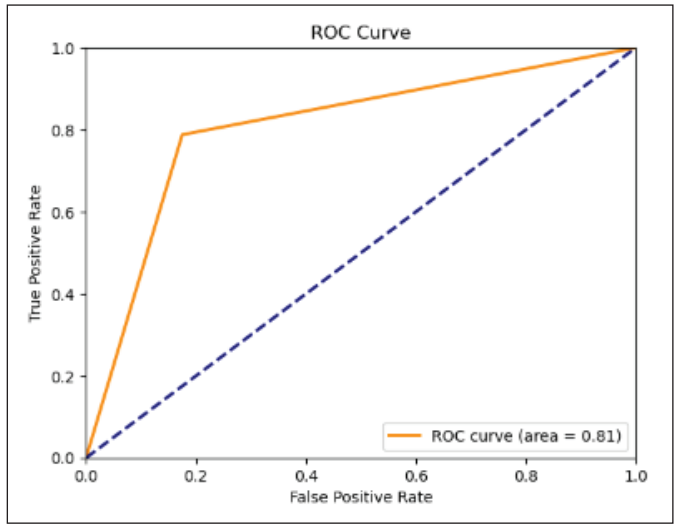

Figure 6. ROC curve.

Table 6. Confusion matrix (Percentage).

		Actual	
		IDC+	IDC-
Predicted	IDC+	98.5	1.5
	IDC-	1.6	98.4

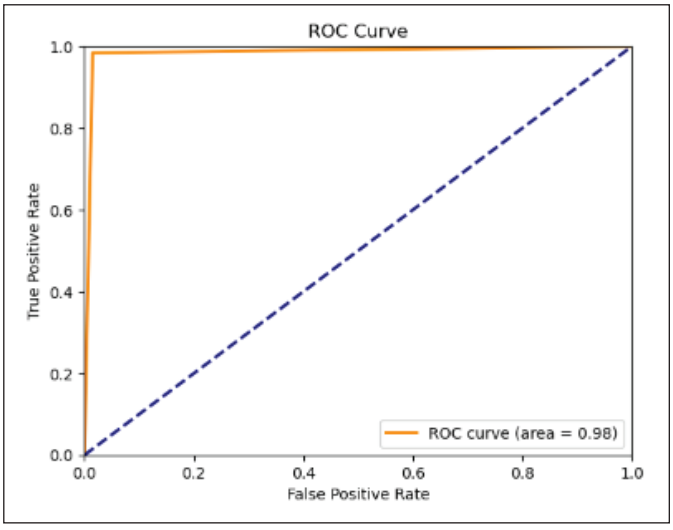

Figure 7. ROC curve.

The lower metric results obtained from Real/Real classifications compared to Synthetic/Synthetic classifications in both cases of ResNet18 classification can be explained as follows: The real dataset contains out of distribution examples within itself, which negatively affect synthetic data generation and classification results. Upon examination of the dataset, many out of distribution data points were found.

The lower accuracy, precision, recall, and F1 score values obtained when using synthetic data for training and real data for testing (Synthetic/Real) compared to Real/Real classifications by 8% can be interpreted as follows:

- As mentioned earlier, the presence of out of distribution examples within the real dataset and the use of real data as test data affect the classification results. The model could not find correlation for out of distribution data.
- The lesser diversity of synthetic data compared to real data might imply that synthetic data have less diversity than real data (since the area learned from the real data distribution during synthetic data generation is small compared to the total distribution of real data, the diversity within synthetic data is low). When synthetic data is tested with real data, low metric values are obtained due to the low diversity of the training data.

Table 7. Confusion matrix (Percentage).

		Actual	
		IDC+	IDC-
Predicted	IDC+	72.6	22.4
	IDC-	23.4	76.6

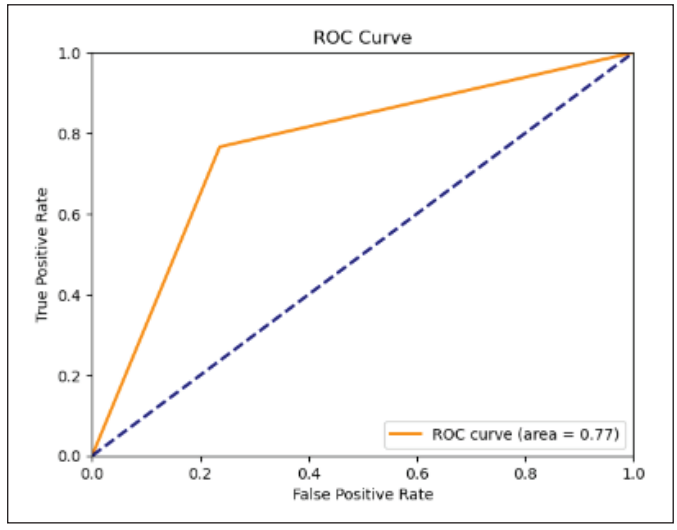

Figure 8. ROC curve.

Figure 8 shows the ROC curve of the classification and Table 7 shows the confusion matrix of the classification.

The similar metric values obtained from Synthetic/Real and Real/Synthetic classifications using the ResNet18 model highlight the similarity between real and synthetic data. Furthermore, achieving high results using a dataset containing out of distribution data demonstrates the success of the model.

The MSG-GAN model used in the study was successful in the production of synthetic data and high accuracy rates were achieved in the classification of these data.

In this study, unlike other studies in the literature, four different classification results that were expected to be related and close to each other were found. In this way, it was tried to determine whether synthetic data could be used instead of real data.

In summary, the results of the classification in four different scenarios were examined to see whether the synthetic data reflected the real world data, and the classification results were found to be related and close to each other. In this case, it has been determined that high quality and similar to reality synthetic data is produced.

Moreover, to evaluate the similarities between synthetic and real image data generated by GANs, various criteria are used. Statistical metrics (mean, variance, distribution, and correlation), machine learning model performance comparisons, visualization techniques (PCA and t-SNE), and feature maps are among these evaluation methods. Ensuring the reliability of these evaluations involves using multiple criteria together, having sufficient data size, ensuring the repeatability of the methods, and performing independent validations. These approaches are crucial for determining how closely GAN-generated synthetic data resembles real data and its reliability in specific applications.

4. Conclusion and Suggestions

This article discusses the increasing importance of data in today's world and the growing need for more data in the technological age, despite the current insufficiency of data in certain areas, particularly in healthcare, due to limitations such as data scarcity, data imbalance, data accessibility, and privacy. Therefore, synthetic data is required. Synthetic data can enrich and diversify real datasets. This enables the model to be trained from a broader perspective and adapt to various conditions. As a solution to the problem, this article proposes generating highly realistic synthetic data from breast cancer histopathology images using MSG-GAN and then classifying both real and synthetic data using the pre-trained ResNet18 model to determine the similarity of synthetic data to real data. Additionally, the article discusses algorithms and architectures from Convolutional Neural Network (CNN) and DL.

This research presents the results of experiments conducted on four different classification tasks to evaluate how close synthetic data is to real data. The results show that synthetic data can perform similarly to real data but must be carefully evaluated. Successful outcomes include the generation of realistic synthetic images, eliminating the need for manual visual breast cancer detection. It is predicted that the production of synthetic data in healthcare will reduce the need for human intervention in the disease detection and diagnosis process and accelerate research in the healthcare field. Therefore, it is concluded that the use of GANs in healthcare for purposes such as data augmentation, dataset enrichment, and balancing is highly beneficial and reliable.

In future studies, the focus should be on improving the process of synthetic data generation using datasets with less skewed distributions and enhancing the quality of synthetic data. Additionally, it is recommended to employ different metrics to further measure the similarity of synthetic data to real data.

5. References

Cruz-Roa, A., Basavanhally, A., González, FA., Gilmore, H., Feldman, MD., Ganesan, S., Shih, N., Tomaszewski, JE., Madabhushi, A. 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. Medical Imaging, Doi: 10.1117/12.2043872

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 289-293. Doi: 10.1109/ISBI.2018.8363576

Goodfellow, IJ., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, AC., Bengio, Y. 2014. Generative adversarial networks. Communications of the ACM, 63(10): 139-144. Doi: 10.1145/3422622

Graves, A. 2013. Generating sequences with recurrent neural networks. ArXiv preprint, abs/1308.0850.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. 2017. Improved training of wasserstein gans. arXiv preprint, arXiv:1704.00028.

He, K., Zhang, X., Ren, S., Sun, J. 2015. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. Doi: 10.1109/CVPR.2016.90

Janowczyk, A., Madabhushi, A. 2016. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform., 7(29). Doi: 10.4103/2153-3539.186902

Karnewar, A., Wang, O. 2019. MSG-GAN: Multi-scale gradients for generative adversarial networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7796-7805. Doi: 10.1109/CVPR42600.2020.00778

Karras, T., Laine, S., Aila, T. 2019. A style-based generator architecture for generative adversarial networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401-4410. Doi: 10.1109/CVPR.2019.00453

Kingma, DP., Ba, J. 2015. Adam: A method for stochastic optimization. CoRR, abs/1412.6980.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324. Doi: 10.1109/5.726791

Liu, S., Setio, AAA., Ghesu, FC., Gibson, E., Grbic, S., Georgescu, B., Comaniciu, D. 2020. No surprises: Training robust lung nodule detection for low-dose CT scans by augmenting with adversarial attacks. IEEE Trans. Med. Imaging, 40(1): 335-345. Doi: 10.1109/TMI.2020.3028311

Salehinejad, H., Colak, E., Dowdell, T., Barfett, J., Valaee, S. 2018. Synthesizing chest X-ray pathology for training deep convolutional neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 990-994. Doi: 10.1109/ICASSP.2018.8461382

Teh, Y., Tan, GH., Taib, NA., Rahmat, K., Westerhout, CJ., Fadzli, F., See, MH., Jamaris, S., Yip, CH. 2015. Opportunistic mammography screening provides effective detection rates in a limited resource healthcare system. BMC Cancer, 15: 405. Doi: 10.1186/s12885-015-1420-4

Xu, B., Wang, N., Chen, T., Li, M. 2015. Empirical evaluation of rectified activations in convolutional networks. ArXiv preprint, abs/1505.00853.

Zhu, JY., Park, T., Isola, P., Efros, AA. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. IEEE International Conference on Computer Vision (ICCV), pp. 2242-2251. Doi: 10.1109/ICCV.2017.244