

Article

Not peer-reviewed version

Enhancing Histopathological Image Classification Performance through Synthetic Data Generation with Generative Adversarial Networks

[José L. Ruiz-Casado](#) , [Miguel A. Molina-Cabello](#) , [Rafael M. Luque-Baena](#) *

Posted Date: 8 May 2024

doi: 10.20944/preprints202405.0467.v1

Keywords: generative adversarial networks; data augmentation; classification; histopathological images; breast cancer



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Enhancing Histopathological Image Classification Performance through Synthetic Data Generation with Generative Adversarial Networks

Jose L. Ruiz-Casado ¹, Miguel A. Molina-Cabello ^{1,2}  and Rafael M. Luque-Baena ^{1,2,*} 

¹ ITIS Software. University of Málaga. C/ Arquitecto Francisco Peñalosa, 18. 29010 Málaga, Spain

² Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND, Málaga, Spain

* Correspondence: rmluque@uma.es

Abstract: Breast cancer is the second most common cancer worldwide, primarily affecting women, while histopathological image analysis is one of the possibilities used to determine tumor malignancy. Regarding image analysis, the application of deep learning has become increasingly prevalent in recent years. However, a significant issue is the unbalanced nature of available datasets, with some classes having more images than others, which may impact the performance of the models due to poorer generalisability. A possible strategy to avoid this problem is downsampling the class with the most images to create a balanced dataset. Nevertheless, this approach is not recommended for small datasets as it can lead to poor model performance. Instead, techniques such as data augmentation are traditionally used to address this issue. These techniques apply simple transformations, such as translation or rotation, to the images to increase variability in the dataset. Recently, Generative Adversarial Networks (GANs) have emerged, which can generate images from a relatively small training set. This work aims to enhance the model's performance in classifying histopathological images by applying data augmentation using GANs instead of traditional techniques.

Keywords: generative adversarial networks; data augmentation; classification; histopathological images; breast cancer

1. Introduction

Early detection of breast cancer represents a crucial strategy for reducing mortality rates. Screening methods, such as mammography and ultrasound, are commonly employed for this purpose. Despite the efficacy of these methods in identifying potential abnormalities, determining the malignancy level of a tumor accurately requires the analysis of histopathological images [1]. This process involves digitization of tissue samples obtained through biopsy, which is an invasive procedure. Manual analysis of histopathological images presents several limitations, including laboriousness and time-consuming nature, in addition to a potential for human error. To address these challenges, Computer-Aided Diagnosis (CAD) [2] systems have been developed to assist healthcare professionals in the decision-making process, providing more accurate and efficient diagnoses.

Technological advances and the development of informatics applications and tools tailored specifically for the processing and analysis of histopathology images have been catalyzed by advancements in technology. These tools have undergone considerable evolution, with modern iterations incorporating cutting-edge deep learning models [3] to enhance their capabilities for image analysis. However, in order for these deep learning neural networks to function successfully, there is a necessity for access to sufficiently large datasets. These datasets are crucial for the training of the neural networks to identify and classify the key features of each class represented in the dataset accurately, and therefore, they must be sufficiently representative.

The acquisition of such datasets presents a significant challenge, primarily due to the high costs and potential risks associated with the procedures used to obtain histopathological images. Furthermore, there is often an inherent imbalance in these datasets, with negative classes being overrepresented compared to positive cases. Addressing these challenges is essential to ensure the

effectiveness and reliability of deep learning models in assisting healthcare professionals with breast cancer diagnosis and treatment planning [4].

The assembly of a comprehensive dataset of breast cancer images represents a crucial step in advancing research and diagnosis in the field. However, the scarcity and imbalance of such datasets pose significant obstacles, hindering the development of accurate and robust machine learning models for breast cancer detection and classification. In this context, Generative Adversarial Networks (GANs) emerge as a powerful tool to overcome dataset limitations by generating synthetic images that complement the available data [5]. By capitalizing on the capacity of GANs to generate realistic and diverse images, researchers can enhance existing datasets, balance class distributions, and address gaps in data coverage. This symbiotic relationship between dataset curation and GAN-based augmentation not only enhances the quality and representativeness of the dataset but also strengthens the performance and generalization capabilities of machine learning models trained on it. Consequently, the integration of GANs into the dataset preparation process represents a promising avenue for addressing the data scarcity inherent in breast cancer imaging datasets and advancing the state-of-the-art in breast cancer detection and diagnosis.

Since the introduction of GANs in the seminal work by Goodfellow et al. ([6]), this architectural paradigm has been widely applied across diverse domains. GANs operate on the principle of employing two distinct models: the generator and the discriminator, engaged in a competitive interplay akin to a zero-sum game. The training of GANs involves fostering competition between two components, namely the generator and the discriminator, which are set up in opposition to one another.

To address the challenges of stability and variability encountered during GAN training, Tero Karras proposed a significant evolution in the network's architecture, as outlined in reference [7]. The fundamental concept introduced involves incrementally enhancing the quality of the generated images. To circumvent the issue of the generator becoming ensnared by overly similar examples, Karras et al. suggest a strategy involving the calculation of the standard deviation across all pixels within the generated mini-batch. This approach enables the discriminator to discern a generated image based on the pixel-level standard deviation present in the batch.

Spectral Normalization Generative Adversarial Networks (SN-GAN) [8] represent a variant of Generative Adversarial Networks (GANs) designed to enhance training stability and improve the quality of generated images. This model addresses the problem of Lipschitz continuity, which is crucial for ensuring convergence and preventing mode collapse in GAN training. SN-GAN achieves this by applying spectral normalization to the weights of the discriminator network, effectively bounding the Lipschitz constant of the discriminator's mapping function. This normalization technique stabilizes the training process by controlling the Lipschitz norm of the discriminator, thereby mitigating the risk of exploding gradients and facilitating smoother optimization. Consequently, SN-GANs are capable of generating high-quality and diverse synthetic images across a range of datasets, making them a valuable tool in the field of generative modeling.

The boundary equilibrium generative adversarial network (BEGAN) [9] represents an original approach to the GAN structure introduced by Berthelot et al. BEGANs seek to overcome the challenge of training stability and image quality by introducing a novel loss function inspired by the concept of equilibria in game theory. In contrast to traditional GANs, which rely on adversarial loss to guide the training process, BEGANs incorporate an additional measure known as the reconstruction loss. This loss is computed by comparing the original input images with their reconstructions generated by the generator network. BEGANs strive to achieve a balance between the adversarial and reconstruction losses, seeking an equilibrium point where the discriminator cannot distinguish between real and generated images, while the generator can faithfully reconstruct the input distribution. By maintaining this equilibrium, BEGANs produce high-quality and visually coherent images, exhibiting improved training stability and mitigating issues such as mode collapse. This innovative architecture has attracted attention for its ability to generate diverse and realistic images across various datasets, contributing to advancements in the field of generative modeling. In order to prevent model collapse, the model

employs techniques proposed in Energy-based generative networks (EBGAN) and Advances in Neural Information Processing Systems. [10].

Re-GAN [11] addresses the challenge of training high-fidelity GANs with limited data. It is designed to alleviate the need for extensive training datasets typically required for such tasks. The method introduces a dynamic architectural reconfiguration process during training, applying pruning to configurations that are not promising [12]. This enables the exploration of diverse sub-network structures within GANs. Re-GAN achieves this by iteratively pruning and regrowing connections, which optimises the network complexity of the GANs. This approach stabilises GAN models, offering an alternative to traditional GAN training methods and the costly process of finding GAN tickets. Furthermore, the versatility of Re-GAN is demonstrated through its capacity to achieve stability across datasets of varying sizes, domains, and resolutions. Additionally, it is able to function effectively with different GAN architectures, thus providing a data-efficient alternative to other GAN training methods. Furthermore, Re-GAN enhances performance when combined with recent augmentation techniques, all while requiring fewer floating-point operations and less training time. Notably, Re-GAN surpasses state-of-the-art StyleGAN2 without the need for additional fine-tuning steps, thereby demonstrating its effectiveness in improving GAN training efficiency and output quality.

To evaluate the performance of generative networks accurately, it is essential to utilize specific loss functions that are tailored to their unique characteristics. The Wasserstein loss function, as introduced by Arjovsky et al. in their seminal work, 'Wasserstein Generative Adversarial Networks' [13], leverages the Wasserstein distance to gauge the quality of generated images, thereby enhancing training stability. It is of paramount importance to adhere to the Lipschitz constraint in order for this loss function to function properly. When a gradient penalty is incorporated to enforce compliance with this constraint, the resulting approach is commonly referred to as the Wasserstein loss function with gradient penalty [14].

To effectively assess a generative model's performance, it is of paramount importance to conduct a comparison between the generated images and a set of exemplar images. This comparison necessitates a degree of quantifiability, which is essential in enabling meaningful comparisons between different generative models, and thus facilitating the identification of the most superior performing model. Among the various metrics that have been introduced in this field, the Frechet Inception Distance (FID), which was first proposed as an objective standard for evaluating generative models by Heusel and colleagues, has proven to be a particularly valuable tool in this regard (Heusel et al. cited in [15]). The use of a dataset that differs from ImageNET [16], which is commonly used as a benchmark, results in higher FID values due to differences in feature vector representations [17,18]. This occurs because of the divergence in image content, style, and distribution between datasets, which amplifies the perceptual dissimilarity captured by the FID metric.

2. Methods

This section outlines the methodology and procedure used to carry out the experiments. This section presents the methodology and procedural steps employed in the conducted experiments. Multiple GANs were trained to be utilized in image enhancement applications to the dataset. Image enhancement encompasses the addition of supplementary images to the original data set. Typically, scalar transformations are employed to enhance the variability of the data set. The trained GANs are utilized to investigate the impact of these networks on the augmentation of images.

In order to accurately assess the impact of generative networks, multiple convolutional models will be utilized, training on the primary image set and on the sets generated with these generative networks. The selected convolutional networks for this evaluation are VGG16, ResNet, and Inception. The selection of these architectures for comparison in the experiments is based on their current status as state-of-the-art convolutional network architectures, which are widely recognized for their efficacy in image recognition tasks. These networks represent distinct approaches to feature extraction and representation learning, allowing for comprehensive evaluation across different architectural

paradigms. Additionally, their availability as pre-trained models with well-established performance benchmarks facilitates fair and rigorous comparison within the scope of the experiments.

The experiments may then be divided into two categories: GAN training and CNN training. For each category, a set of configurations and metrics will be established to enable the results of each experiment to be compared.

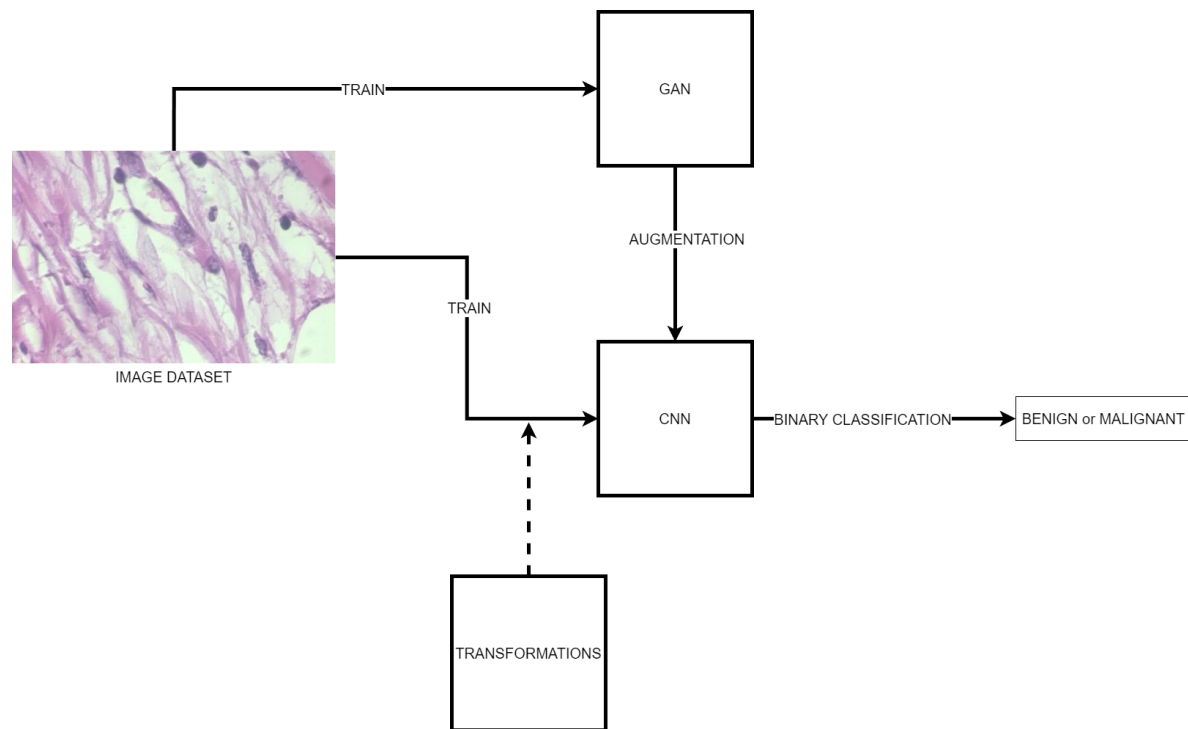


Figure 1. Graphical abstract for conducted experiments.

Regarding the dataset to be used, it is provided by *Laboratório Visão Robótica e Imagem UFPR* [19]. The images are 700x460 in color, which means they have 3 channels. The images are divided into two classes:

- Benign. With 588 images.
- Malignant. With 1233 images.

Figure 2 presents a visual representation of samples from the benign class on the left and the malignant class on the right. This illustration provides a clear visual comparison of histopathological images representing different classes of breast cancer.

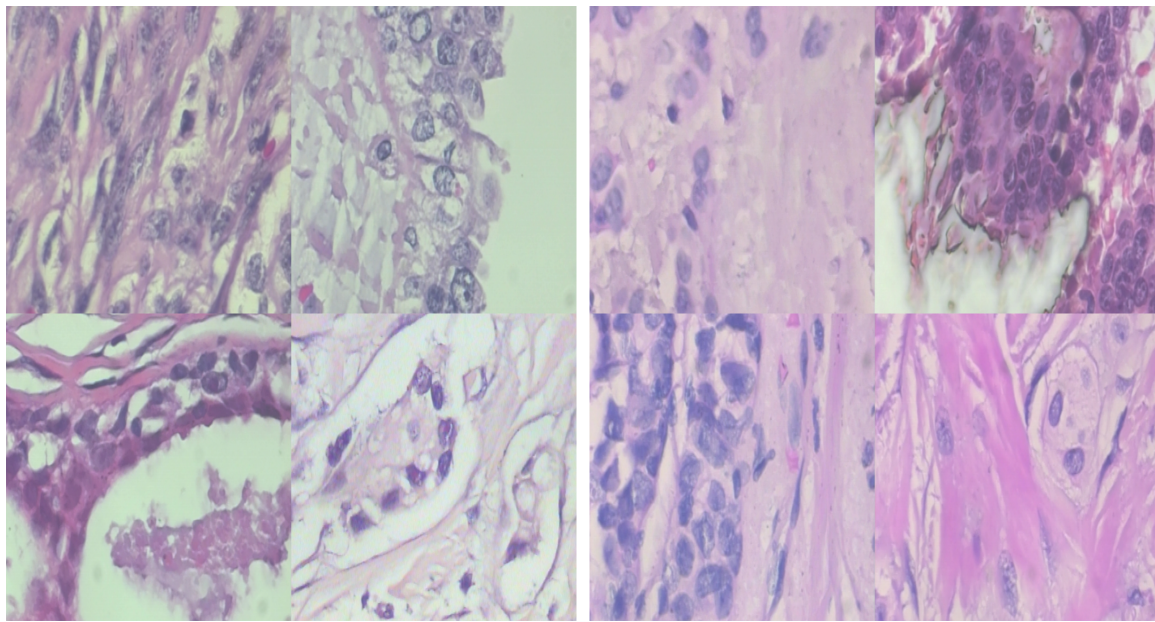


Figure 2. Samples of benign class (left side) and malignant class (right side).

2.1. GAN Training

The issue with the medical imaging set in use is that there are a greater number of malignant images than benign images. In order to address this imbalance in the dataset, several GAN architectures will be trained in an effort to identify the model capable of achieving superior results.

In order to facilitate a comparative analysis of the various GAN architectures, the FID metric will be employed. It should be noted that this metric is not expected to exhibit a specific range of values, given that it is highly dependent on the pre-trained weights utilized and the nature of the images employed. Consequently, the FID value for each architecture can be compared with those of the other architectures, with the lowest value representing the optimal performance. Furthermore, a visual comparison of the generated images will be conducted in order to ascertain the degree of similarity between them and the original images. To compute the metric the package PyTorch FID [20] is used.

The optimizer selected for training GANs is ADAM. This decision is based on its effectiveness in addressing the unique challenges posed by GAN optimization. ADAM is distinguished by its adaptive learning rate capabilities, which enable it to dynamically adjust the learning rate for each parameter based on past gradients and moment estimates. This adaptive behavior is particularly beneficial for GAN training, where the generator and discriminator networks often have different learning dynamics and can benefit from different learning rates. Furthermore, Adam's momentum and second-order moment estimation mechanisms assist in the mitigation of issues such as vanishing or exploding gradients, which are prevalent in GAN training due to the non-linear and highly volatile nature of the optimization landscape. In summary, Adam's adaptive learning rate, momentum, and moment estimation features render it an optimal optimizer for GAN training, enabling more stable and efficient convergence while reducing the necessity for manual tuning of learning rates.

2.1.1. ProGAN

A good set balancing system can be obtained by training two generative models, one for each class, using the ProGAN architecture. It is important to note that both models are trained with the same hyperparameter settings. To choose the size of the generated images, the depth value d must be selected in order to calculate the output size as $2^{(d+1)}$. In consideration of the dimensions of the generated images, a depth value of 7 was selected to ensure that the resulting images are 256x256 pixels in size.

It is necessary to select an appropriate batch size for each depth level, taking into account the size of the output. It is important to note that the size of the selected image must be equal to the size of the output at each level, in order to prevent an erroneous operation in the discriminator and ensure the consistency of the model. The Table 1 shows the selected value at each depth level.

Table 1. Batch size adjustment by depth level.

Depth	1	2	3	4	5	6	7
Batch size	64	64	32	16	4	2	1
Image size	4	8	16	32	64	124	256

When a new block is added when the depth level is changed, it is not added directly, but rather, is faded in with the previous block. The degree to which the new block fades in is determined by a fade percentage variable, which indicates the percentage of the new block that is utilized in the output. The value of this variable is set to 0.5, thereby establishing a baseline for the degree of fade in.

An external parameter that facilitates accelerated training is the number of threads utilized for loading the training set. In this instance, four loading threads have been defined.

2.1.2. BEGAN

This section examines the configuration of Boundary Equilibrium Generative Adversarial Networks (BEGAN) with the specific aim of addressing the imbalance inherent in image datasets. BEGAN, renowned for its ability to generate high-quality and diverse images while maintaining training stability, presents a promising solution for dataset-balancing tasks. By harnessing the equilibrium between generator and discriminator losses, BEGAN offers a principled framework for generating synthetic images that effectively supplement the existing dataset, thereby promoting class balance and enhancing the robustness of machine learning models. This section explores the key considerations and methodologies involved in configuring BEGAN for dataset balancing, offering insights into its application in the realm of image dataset management and augmentation.

A total of 500 epochs are utilized for the training process, providing sufficient iterations for the model to learn and refine its parameters. Each epoch involves the processing of data in batches, with a batch size of 64 samples per iteration, ensuring a balance between computational efficiency and model stability. Additionally, the latent dimension of 128 is chosen, determining the dimensionality of the input noise vector that serves as the basis for generating synthetic images. This configuration ensures that the model explores a sufficiently rich and diverse latent space, thereby facilitating the generation of high-quality and varied images throughout the training process.

The learning rate of 0.0002 strikes a balance between rapid convergence and stability, preventing the model from oscillating or diverging during training. Additionally, the choice of b1 and b2 parameters (0.9 and 0.999, respectively) controls the exponential decay rates of the first and second moments of gradients, influencing the optimizer’s behavior. These values are selected to ensure smooth and consistent optimization, allowing the model to effectively learn the intricate dynamics of the data distribution and maintain equilibrium between the generator and discriminator networks throughout training. Overall, the Adam optimizer, with its tailored hyperparameters, provides an effective framework for training BEGAN models, facilitating the generation of high-quality images while promoting training stability and convergence.

2.1.3. SNGAN

In accordance with the previously outlined training configurations, a total of 500 epochs are selected to provide sufficient iterations for the model to learn and refine its parameters. Furthermore, the selection of a latent vector size of 512 contributes to the model’s expressive capacity, enabling it to capture a more diverse range of latent representations. This configuration strikes a balance between model complexity and computational efficiency, facilitating robust and high-quality image

generation across various datasets and tasks. In summary, the simplicity and effectiveness of the SNGAN configuration demonstrate its versatility and efficacy in the field of generative modeling.

2.1.4. REGAN

In order to assess the effectiveness of ReGAN across different generative modeling paradigms, the chosen architectures are ProGAN and SNGAN. These architectures have been selected as they possess distinct characteristics and capabilities that will enable the assessment to be carried out. In alignment with the established training protocols for ProGAN and SNGAN, the same configuration parameters are employed during ReGAN training. This ensures consistency and comparability across experiments. The objective of this study is to elucidate the impact of ReGAN within the framework of ProGAN and SNGAN with identical configurations on training dynamics, convergence, and the quality of generated outputs. This will provide valuable insights into the effectiveness of ReGAN across different GAN architectures.

2.2. CNN Training

This section of experiments is a comprehensive analysis of convolutional neural network (CNN) training using various datasets and augmentation techniques. The objective is to evaluate the impact of dataset augmentation, traditional transformations, and dataset balancing with four different GAN architectures (ProGAN, SNGAN, BEGAN, and ReGAN) on the performance of CNN models. Three different CNN architectures, namely VGG16, Inception, and ResNet, are evaluated in this experiment. These networks have proven to be effective in image classification tasks. To investigate the impact of dataset augmentation and GAN-based balancing on CNN performance, we train these architectures on three datasets: an original dataset, an augmented dataset using traditional transformations, and a balanced dataset generated with each GAN. The objective of this experimental endeavor is to provide valuable insights into the efficacy of data augmentation and GAN-based data balancing techniques in enhancing the performance and robustness of CNNs. This, in turn, will contribute to advancements in image classification and machine learning research.

In order to ensure the robustness and reliability of the performance metrics, the KFold cross-validation technique has been employed to partition the training dataset consistently at each time point throughout the experimentation process. With a choice of $k = 5$, the dataset is divided into five equally sized folds, with each fold serving as a validation set while the remaining folds are utilized for training. This iterative process is repeated five times, with each fold taking a turn as the validation set. By systematically rotating the dataset partitions, KFold cross-validation minimizes the risk of overfitting and provides a more accurate estimation of model performance across different subsets of the data. This approach enhances the generalization capabilities of the trained models, ensuring that the reported performance metrics are representative and reliable. Through the meticulous application of KFold cross-validation, comprehensive insights into the robustness and effectiveness of the CNN models trained on various datasets and augmented with different techniques are obtained.

Given the limited number of hyperparameters inherent in these networks, their configurations are the same to avoid bias in the results. Therefore, all networks will be trained for the full 500 epochs to ensure thorough learning and optimization. In addition, a uniform batch size of 32 will be used across all experiments to maintain consistency and comparability in the training process. By standardizing these key training parameters, we aim to eliminate potential sources of variation and ensure the reliability and fairness of experimental results across different CNN architectures and datasets.

Cross Entropy was chosen as the loss function because of its effectiveness in classification tasks. The Cross-Entropy loss measures the discrepancy between the predicted class probabilities and the true class labels, penalizing the model for incorrect predictions while encouraging accurate classification. This loss function is well suited for training Convolutional Neural Networks (CNNs) as it efficiently captures the model's ability to discriminate between different classes and provides meaningful gradients for optimization. By minimizing the cross-entropy loss during training, CNNs are encouraged

to learn discriminative features and make more accurate predictions on unseen data. This choice of the loss function is consistent with the goal of achieving optimal classification performance and underscores the emphasis on accuracy and generalization in the experimental setup.

In addition to selecting Cross Entropy as the loss function, the Adam optimizer was chosen to facilitate the training process. In this configuration, Adam operates at a learning rate of 0.0002, which balances fast convergence with stability, allowing for smooth and efficient training. In addition, the choice of betas, set at 0.5 and 0.999, determines the exponential decay rates for the first and second moments of the gradient, respectively. By adjusting these parameters, Adam optimizes the balance between exploration and exploitation during training, allowing the CNNs to effectively navigate the optimization landscape and converge to an optimal solution. This configuration of the Adam optimizer underscores the commitment to achieving robust and reliable results in the experimental setup and highlights the importance of optimization techniques in training deep learning models for image classification tasks.

3. Results

This section presents the outcomes of two distinct experiments conducted to evaluate the efficacy of GAN training and CNN training with balanced datasets. The first experiment focused on utilizing GANs, specifically ProGAN, SNGAN, BEGAN, and ReGAN, to balance the dataset. This was done with the goal of mitigating class imbalance and enhancing the diversity of the training data. The impact of GAN-based dataset balancing techniques on the performance of CNNs trained on these balanced datasets is examined in this experiment. In the second experiment, CNNs are trained directly on the balanced dataset without GAN augmentation, serving as a benchmark to compare against the GAN-trained models. By contrasting the outcomes of these two approaches, insights are gained into the effectiveness of GAN-based dataset balancing in improving CNN performance and robustness in image classification tasks.

3.1. GAN Training

In order to conduct a comprehensive comparison of the various GAN architectures, the FID metric is employed. The values of FID rely on the pre-trained weights of the Inception network, as highlighted in references [17,18]. It is imperative to utilize the same Inception network across all GAN architectures to ensure consistency and fairness in the comparison process. The utilization of a standardized Inception network ensures uniformity within the evaluation framework, thereby enabling an unbiased assessment of the performance of each GAN architecture. This approach allows for meaningful comparisons based on the FID metric, facilitating insights into the relative strengths and weaknesses of the different GAN models.

Table 2. FID values achieved by each GAN architecture. Best result is highlighted in bold.

Model	Benign	Malignant
BEGAN	349.09	331.16
PROGAN	261.51	222.81
SNGAN	424.93	401.71
REGAN PROGAN	372.64	306.53
REGAN SNGAN	473.95	485.67

Upon examination of the FID results, it becomes evident that all models, except REGAN SNGAN, exhibit FID values below 400. This suggests a relatively favorable performance in terms of image quality and diversity. In particular, PROGAN presents the lowest FID values for the benign and malignant classes, indicating a superior performance in generating high-quality images compared to the other models. However, the FID values for REGAN SNGAN exceed the threshold, indicating

potential challenges or limitations in achieving satisfactory quality in image generation. These FID results provide valuable information on the performance of the various models, guiding future analysis and refinement efforts.

GANs demonstrating FID values below 400 are deemed suitable candidates for dataset balancing. This criterion ensures that the chosen GANs exhibit favorable performance in generating high-quality and diverse images, thereby effectively balancing the dataset. Furthermore, it is crucial to highlight that the outputs generated by the selected GANs seamlessly adapt to the input requirements of the CNNs used for image classification. By aligning the characteristics of the GANs' outputs with the specifications of the CNNs' inputs, a smooth integration is achieved, facilitating the utilization of GAN-generated data to enhance the robustness and performance of the CNN models in image classification tasks. This symbiotic relationship between GANs and CNNs underscores their complementary roles in the dataset balancing and classification processes, ultimately contributing to improved model accuracy and generalization capabilities.

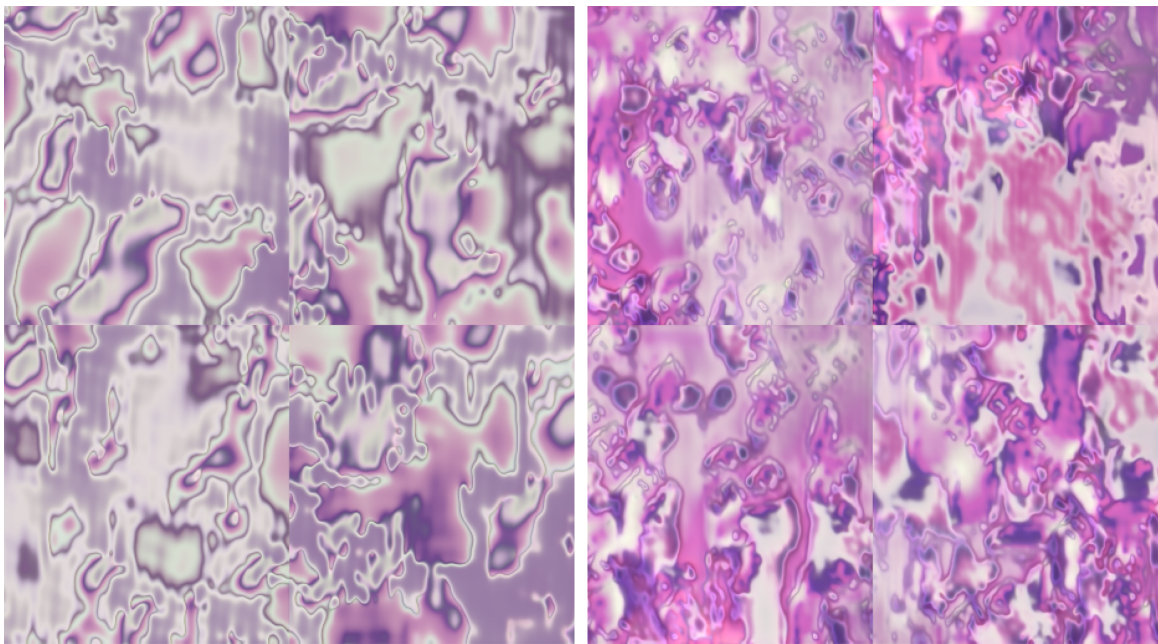


Figure 3. Images generated for benign class (left) and malignant class (right) with ProGAN, which is the GAN that achieves the lowest FID value in our experiments.

The presented graphs offer a comprehensive overview of the performance dynamics of CNNs across various epochs and folds in a five-fold cross-validation setup. Each graph provides insights into the mean accuracy and standard deviation of CNN models trained using different datasets and methodologies. The upper side graph illustrates the performance metrics derived from training the CNN models on the original dataset. In contrast, the lower side graph showcases the corresponding metrics obtained from each data augmentation technique applied to the dataset (GAN). In both graphs, three distinct CNN architectures (Inception, ResNet, and VGG16) are evaluated, with mean accuracy and standard deviation depicted for each architecture. The shaded regions surrounding the lines represent the standard deviation, offering insights into the variability of model performance across different folds and epochs. These visualizations provide a comprehensive understanding of the impact of data augmentation techniques on CNN training and performance variability, informing further analysis and optimization efforts in image classification tasks.

A comparison of the two graphs in Figure 4 reveals an improvement in mean accuracy when traditional data augmentation techniques are applied to the dataset. However, despite this enhancement in accuracy, it is evident that the standard deviation remains relatively high across epochs and folds.

This observation suggests that while traditional data augmentation contributes to improved overall performance, it may not fully address the variability in model performance across different folds and epochs. Consequently, while mean accuracy is positively influenced by traditional transformation, the persistently high standard deviation highlights the necessity for further investigation and refinement in training methodologies to achieve greater consistency and stability in CNN performance.



Figure 4. Comparison of Training Data Performance: Original dataset (upper) vs. Traditional augmentation techniques (lower).

Figure 5 presents a comparison graph illustrating the training mean accuracy and standard deviation using BEGAN as the dataset balancer. The graph reveals a notable improvement in mean accuracy across all three CNN architectures, indicating enhanced model performance when trained on the dataset balanced with BEGAN. Additionally, there is an overall reduction in standard deviation, suggesting improved stability and consistency in model performance. Nevertheless, it is notable that the standard deviation remains relatively high for the Inception architecture, indicating variability in performance across different epochs and folds. Despite this exception, the general trend reflects the efficacy of BEGAN in improving both accuracy and stability for the CNN models, underscoring its potential as a dataset-balancing technique in image classification tasks.

Figure 6 presents a comparison between CNNs trained on the original dataset (upper side) and CNNs trained with ProGAN as the dataset balancer (lower side). While the accuracy demonstrates improvement with ProGAN-balanced datasets, the higher standard deviation indicates a noticeable increase in instability. Despite ProGAN having the lowest FID among the GAN architectures considered, the results suggest a trade-off between image quality and variability. The CNNs trained with ProGAN-balanced datasets exhibit high-quality images, as reflected by the low FID values. However, they also display more significant variability in performance across epochs and folds, as evidenced by the elevated standard deviation. These findings underscore the complex interplay between image quality and stability in CNN training, highlighting the need for further exploration and optimization to balance these competing factors.

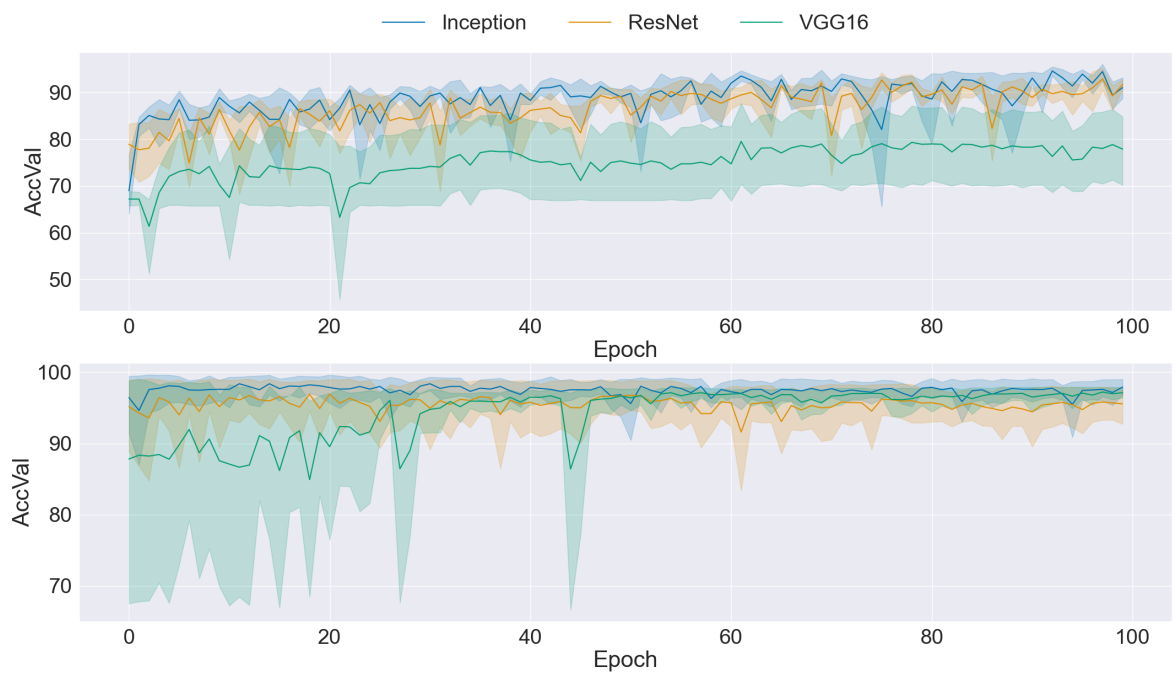


Figure 5. Comparison of Training Data Performance: Original dataset (upper) vs. BEGAN augmentation (lower).

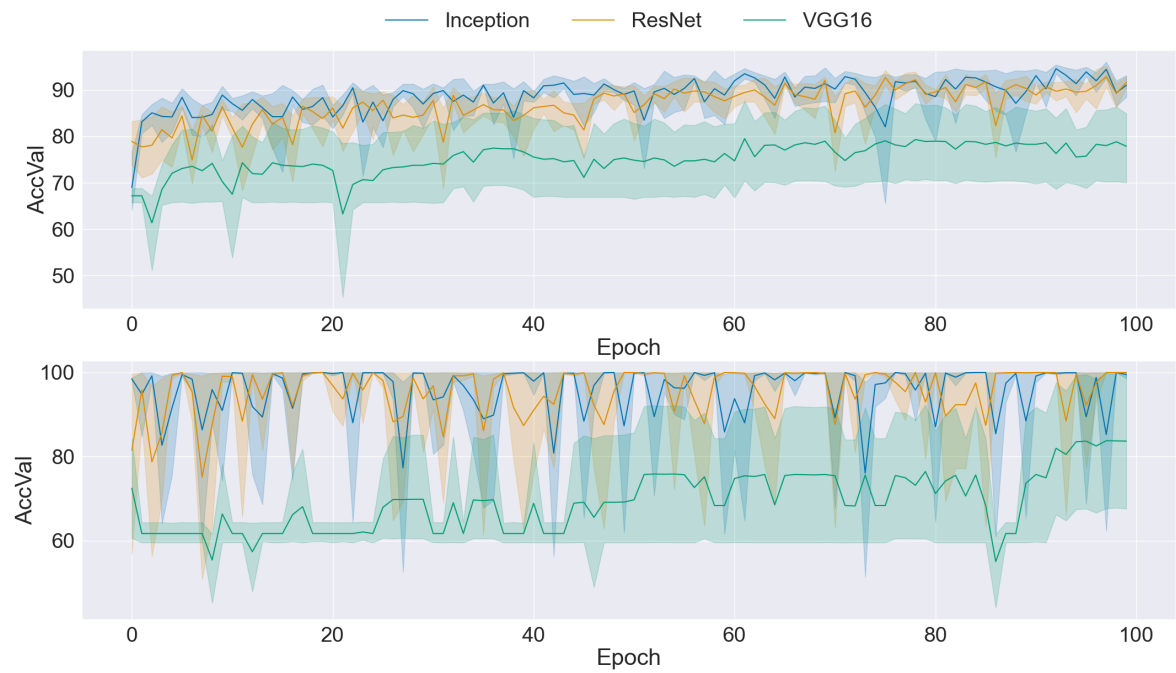


Figure 6. Comparison of Training Data Performance: Original dataset (upper) vs. ProGAN augmentation (lower).

Figure 7 presents a comparison between CNNs trained with original and REGAN ProGAN balanced datasets. While the accuracy achieved with REGAN ProGAN is not as high as with ProGAN, there is a notable improvement in stability, as evidenced by the reduced variability in accuracy values across epochs and folds. However, despite this improvement in stability, the standard deviation remains high, indicating ongoing variability in model performance. Although REGAN ProGAN demonstrates a balance between accuracy and stability, the persistence of a high standard devia-

tion indicates that further optimization may be necessary to achieve greater consistency in model performance.

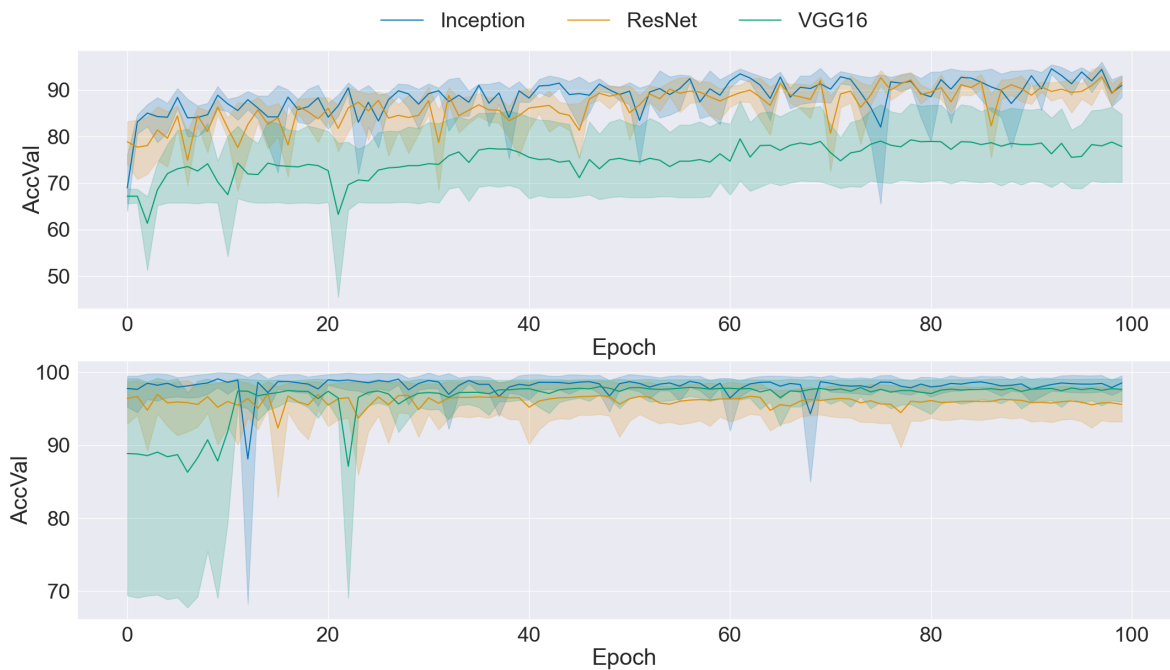


Figure 7. Comparison of Training Data Performance: Original dataset (upper) vs. REGAN ProGAN augmentation (lower).

Table 3 presents the mean accuracy and mean loss, alongside their respective standard deviations, for various CNN models trained with different data augmentation techniques. The table illustrates the impact of distinct augmentation strategies on model performance. Traditional image augmentation demonstrates a modest enhancement in mean accuracy, coupled with a reduction in standard deviation, implying improved consistency in model performance. Conversely, augmentation techniques involving generative adversarial networks (GANs) exhibit significant improvements in both mean accuracy and standard deviation, indicating enhanced model robustness and stability. It is noteworthy that REGAN ProGAN emerges as the top performer, exhibiting the highest mean accuracy and the lowest standard deviation among the evaluated methods.

The outcomes of the Inception model indicate that conventional image transformations result in a marginal improvement in both accuracy and standard deviation. Conversely, GAN-based augmentation techniques demonstrate a substantial enhancement in both accuracy and standard deviation, thereby underscoring their efficacy in improving model performance. Among the GAN-based augmentation methods, ProGAN stands out with the highest accuracy, while maintaining a similar standard deviation compared to other techniques.

The ResNet model presents analogous results, corroborating the trends observed for the Inception model. In all cases, GAN-based augmentation techniques consistently provide the highest average accuracy compared to the original dataset and traditional augmentation methods. It should be noted that although GAN-based augmentation improves accuracy, it is often accompanied by a slightly higher standard deviation, indicating greater variability in model performance. Nevertheless, the REGAN ProGAN model (highlighted in the table) exhibits a notable improvement in accuracy, accompanied by a marginal increase in std.

The VGG16 model emerges as the lowest performing CNN among the architectural variants considered. Overall, VGG16 has a lower average accuracy and a higher standard deviation compared to the other CNN models. Despite these challenges, it can be observed that remarkable performance improvements are achieved with the REGAN ProGAN and BEGAN enhancement techniques. Despite

the inherent limitations of VGG16, REGAN ProGAN (highlighted in the table as the best performer) and to a lesser extent BEGAN demonstrate commendable performance, with higher mean accuracy and relatively lower standard deviation compared to other augmentation methods. These results underscore the resilience and effectiveness of REGAN ProGAN and BEGAN in improving the performance of even the worst performing CNN architecture.

Table 3. Test results for CNNs training. Best values for each model are highlighted in **bold**.

Model	Dataset	Acc	Loss
Inception	Original	92.03 ± 3.64	0.23 ± 0.13
Inception	Traditional Augmentation	92.41 ± 2.81	0.19 ± 0.05
Inception	BEGAN	97.97 ± 1.54	0.07 ± 0.06
Inception	ProGAN	99.36 ± 2.57	0.02 ± 0.04
Inception	REGAN ProGAN	98.45 ± 1.46	0.06 ± 0.06
ResNET	Original	92.63 ± 2.28	0.21 ± 0.06
ResNET	Traditional Augmentation	90.98 ± 4.36	0.23 ± 0.10
ResNET	BEGAN	95.97 ± 4.13	0.22 ± 0.26
ResNET	ProGAN	97.26 ± 5.01	0.01 ± 0.14
ResNET	REGAN ProGAN	96.47 ± 2.72	0.16 ± 0.16
VGG16	Original	79.45 ± 10.64	0.47 ± 0.14
VGG16	Traditional Augmentation	78.57 ± 8.79	0.48 ± 0.14
VGG16	BEGAN	96.52 ± 5.43	0.23 ± 0.08
VGG16	ProGAN	85.10 ± 20.21	0.26 ± 0.35
VGG16	REGAN ProGAN	94.47 ± 2.10	0.11 ± 0.10

4. Conclusions

The research direction established for this study is to leverage generative networks (GANs) to enhance breast cancer diagnosis systems based on histopathological images. To this end, a variety of generative architectures (BEGAN, ProGAN, SNGAN, and REGAN) were trained and evaluated in order to ascertain their efficacy. Following the training of these generative models, their effectiveness in balancing the existing image datasets has been examined. The study specifically investigates the enhancements resulting from incorporating these balanced datasets into the training of state-of-the-art convolutional networks (CNN), including Inception, ResNet, and VGG16.

Upon examination of the FID results, it becomes evident that ProGAN attains the lowest value, indicative of its ability to generate highly realistic images. This highlights the capacity of ProGAN to generate images that closely resemble authentic samples, a pivotal aspect in optimizing the efficacy of breast cancer diagnosis systems. Moreover, both BEGAN and REGAN ProGAN exhibit commendable FID values, which suggests that they are capable of generating images with a high degree of realism. These findings demonstrate the efficacy of BEGAN and REGAN ProGAN as viable alternatives for dataset balancing and image generation. They also validate their utility in improving the performance of breast cancer diagnosis systems based on histopathological images.

The outcomes of training the CNNs demonstrate a notable trend: higher mean accuracy and lower standard deviation (STD) are consistently observed compared to both the original dataset and traditional augmentation methods. This overarching trend demonstrates the efficacy of GAN-based augmentation techniques in enhancing the robustness and stability of the trained models. It is notable that ProGAN, despite exhibiting the lowest FID value, indicative of its ability to generate highly realistic images, exhibits a relatively higher standard deviation. This observation indicates that while ProGAN is highly effective in generating realistic images, it may not explore as much variability in the

generated samples. Nevertheless, the superiority of GAN-based augmentation techniques is evident across all CNN architectures, including those with lower performance, such as VGG16. In conclusion, the findings demonstrate the efficacy of GAN-based augmentation in improving model performance, indicating its potential to enhance breast cancer diagnosis systems based on histopathological images, regardless of the CNN architecture used.

Funding: This work was partially supported by the Autonomous Government of Andalusia (Spain) under grant UMA20-FEDERJA-108, the Ministry of Science and Innovation of Spain under grant PID2022-136764OA-I00, the University of Málaga (Spain) under grants B4-2022, B1-2021_20 and B1-2022_14, and the Fundación Unicaja under grant PUNI-003_2023.

Data Availability Statement: The dataset used in this study is openly available [19]. The source code developed for this work is openly available¹.

Acknowledgments: The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Malaga. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of a RTX A6000 GPU with 48Gb. The authors also thankfully acknowledge the grant of the Universidad de Málaga and the Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gurcan, M.N.; Boucheron, L.E.; Can, A.; Madabhushi, A.; Rajpoot, N.M.; Yener, B. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering* **2009**, *2*, 147–171.
2. Veta, M.; Pluim, J.P.W.; Van Diest, P.J.; Viergever, M.A. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering* **2014**, *61*, 1400 – 1411. <https://doi.org/10.1109/TBME.2014.2303852>.
3. Janowczyk, A.; Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics* **2016**, *7*, 29.
4. Leevy, J.L.; Khoshgoftaar, T.M.; Bauder, R.A.; Seliya, N. A survey on addressing high-class imbalance in big data. *Journal of Big Data* **2018**, *5*, 1–30.
5. Motamed, S.; Rogalla, P.; Khalvati, F. Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. *Informatics in medicine unlocked* **2021**, *27*, 100779.
6. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. <https://doi.org/10.1145/3422622>.
7. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, April 30 - May 3, Conference Track Proceedings, 2018.
8. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* **2018**.
9. Berthelot, D.; Schumm, T.; Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* **2017**.
10. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126* **2016**.
11. Saxena, D.; Cao, J.; Xu, J.; Kulshrestha, T. Re-GAN: Data-Efficient GANs Training via Architectural Reconfiguration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 16230–16240.
12. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *arXiv preprint arXiv:1510.00149* **2015**.

¹ <http://uma.es>

13. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the Proceedings of the 34th International Conference on Machine Learning, 06–11 Aug 2017, Vol. 70, pp. 214–223.
14. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein GANs. In Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, p. 5769–5779.
15. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, p. 6629–6640.
16. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.
17. Kynkäänniemi, T.; Karras, T.; Aittala, M.; Aila, T.; Lehtinen, J. The Role of ImageNet Classes in Fréchet Inception Distance. In Proceedings of the The Eleventh International Conference on Learning Representations, (ICLR), 2023.
18. Pang, T.; Wong, J.H.D.; Ng, W.L.; Chan, C.S. Semi-supervised GAN-based Radiomics Model for Data Augmentation in Breast Ultrasound Mass Classification. *Computer Methods and Programs in Biomedicine* **2021**, *203*, 106018. <https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106018>.
19. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering* **2016**, *63*, 1455 – 1462. <https://doi.org/10.1109/TBME.2015.2496264>.
20. Seitzer, M. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, 2020. Version 0.3.0.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.