applied sciences

MDPI

Article

Data Augmentation in Histopathological Classification: An Analysis Exploring GANs with XAI and Vision Transformers

Guilherme Botazzo Rozendo [1,*], Bianca Lançoni de Oliveira Garcia [1], Vinicius Augusto Toreli Borgue [1], Alessandra Lumini [2], Thaína Aparecida Azevedo Tosta [3], Marcelo Zanchetta do Nascimento [4] and Leandro Alves Neves [1]

[1] Department of Computer Science and Statistics (DCCE), São Paulo State University (UNESP), Rua Cristóvão Colombo, 2265, São José do Rio Preto 15054-000, SP, Brazil; bianca.lanconi@unesp.br (B.L.d.O.G.); vinicius.borgue@unesp.br (V.A.T.B.); leandro.neves@unesp.br (L.A.N.)

[2] Department of Computer Science and Engineering (DISI), University of Bologna, Via dell' Università, 50, 47522 Cesena, Italy; alessandra.lumini@unibo.it

[3] Science and Technology Institute, Federal University of São Paulo (UNIFESP), Avenida Cesare Mansueto Giulio Lattes, 1201, São José dos Campos 12247-014, SP, Brazil; tosta.thaina@gmail.com

[4] Faculty of Computer Science (FACOM), Federal University of Uberlândia (UFU), Avenida João Naves de Ávila, 2121, Bl.B, Uberlândia 38400-902, MG, Brazil; marcelo.zanchetta@gmail.com

* Correspondence: guilherme.botazzo@unesp.br

Abstract: Generative adversarial networks (GANs) create images by pitting a generator (G) against a discriminator (D) network, aiming to find a balance between the networks. However, achieving this balance is difficult because G is trained based on just one value representing D's prediction, and only D can access image features. We introduce a novel approach for training GANs using explainable artificial intelligence (XAI) to enhance the quality and diversity of generated images in histopathological datasets. We leverage XAI to extract feature information from D and incorporate it into G via the loss function, a unique strategy not previously explored in this context. We demonstrate that this approach enriches the training with relevant information and promotes improved quality and more variability in the artificial images, decreasing the FID by up to 32.7% compared to traditional methods. In the data augmentation task, these images improve the classification accuracy of Transformer models by up to 3.81% compared to models without data augmentation and up to 3.01% compared to traditional GAN data augmentation. The Saliency method provides G with the most informative feature information. Overall, our work highlights the potential of XAI for enhancing GAN training and suggests avenues for further exploration in this field.

Keywords: generative adversarial networks; explainable artificial intelligence; GAN training; data augmentation; histopathological classification; vision transformers

check for updates

Citation: Rozendo, G.B.; Garcia, B.L.d.O.; Borgue, V.A.T.; Lumini, A.; Tosta, T.A.; do Nascimento, M.Z.; Neves, L.A. Data Augmentation in Histopathological Classification: An Analysis Exploring GANs with XAI and Vision Transformers. *Appl. Sci.* **2024**, *14*, 8125. https://doi.org/10.3390/app14188125

Academic Editors: Mukul Shirvaikar and Chung Hyun Goh

Received: 30 July 2024
Revised: 30 August 2024
Accepted: 7 September 2024
Published: 10 September 2024

1. Introduction

The performance of machine learning systems heavily relies on data representation. Traditionally, in computer vision, data were represented using handcrafted methods designed to extract specific image features. However, this type of representation requires considerable effort from experts to design and develop techniques that do not always provide the expected return or performance [1]. In recent years, with the increase in the computational power of hardware devices, deep learning algorithms have emerged as a relevant alternative to handcrafted methods. Deep learning automatically transforms raw data, such as pixels from a digital image, into a feature vector, allowing the best data representation to be learned automatically through training. This approach is advantageous because it learns hierarchical representations that capture intricate patterns and structures through multiple layers of abstraction. Also, deep learning models can be fine-tuned for various tasks through transfer learning, making them highly versatile across different domains [1,2].

Training representation learning methods, particularly advanced architectures like neural networks and transformers, can be challenging because they require a large amount of labeled data. Many real application contexts, including the medical context, suffer from low availability of labeled images and adverse conditions, such as class imbalance and a lack of standardization in dimensions and formats [3,4]. These facts compromise adequate training and may cause overfitting. One way to address these problems is by utilizing automatic data generation methods, such as generative adversarial networks (GANs) [5,6]. GANs allow the generation of synthetic images through competitive training between two neural networks, the generator (G) and the discriminator (D) [5]. During training, G tries to generate images that resemble authentic data, while D tries to classify correctly whether the images are original or generated. The adversarial training strategy is based on game theory, in which the objective is to reach the Nash equilibrium [7], where neither G nor D can unilaterally improve their outcomes. This situation occurs when G generates images that are so realistic that D cannot reliably determine if an image is real or fake. At this point, G has learned to produce data that effectively fools the discriminator [8]. Data augmentation through GANs is particularly important in small medical image datasets because it helps to address the issue of overfitting, where models become too tailored to the limited training data and fail to generalize well to new, unseen cases. By generating a wider variety of images, GANs create a more representative training set, which can lead to more robust and accurate models.

However, training GANs is challenging due to issues related to backpropagation, particularly in how G's weights are updated. During backpropagation, the gradients of G are derived from D's predictions, but G has no information on how the images' features contribute to the classification. Because of this, adversarial training tends to be an unbalanced game where D generally has the advantage over G [9,10]. Consequently, D tends to assign higher scores to original images during training, and G fails to fool D even after the model converges [8,9]. Studies in the literature traditionally aimed to improve GAN training by modifying the discriminator. For instance, in DCGAN, the first GAN proposal with convolutional layers, the authors used batch normalization and leaky ReLU activations between D's intermediate layers to make it more stable [11]. The loss function, however, was the Jensen–Shannon divergence [5], which can lead to mode collapse and vanishing gradients [8,11]. The WGAN-GP [12] addresses these problems by using the Wasserstein distance as a loss function, which provides a more meaningful measure of the difference between the probability distributions. WGAN-GP also introduces a gradient penalty term that penalizes the norm of D's gradients, encouraging more diverse generated samples by enforcing a more uniform distribution of gradients throughout the data space. The RAGAN [13] introduces the relativistic discriminator that estimates the probability that an authentic sample is more realistic than a fake sample and vice versa. It considers the relative realness of real and fake samples, providing more informative feedback to G. RAGAN delivers a more nuanced signal to G, allowing it to understand better how to generate plausible samples on its own and the actual data distribution.

Recent techniques have focused on improving the generator's training rather than the discriminator. The approaches involve providing more information about the images' features to G and making the competition between G and D more balanced. For instance, Wang et al. [9] proposed a training technique that raises the spatial awareness of G. The strategy consists of sampling multi-level heatmaps from D using Grad-CAM and integrating them into the feature maps of G via the spatial encoding layer. The authors used D as a regularizer, aligning the spatial awareness of G with D's attention maps. Bai et al. [10] argue that since G's weights are updated only with gradients derived from D, D acts as a referee rather than a player. The authors propose a new training approach with a generator-leading task to make the adversarial game fairer. In this task, D must extract features G can decode to reconstruct the input.

Another possible solution is to combine GANs with explainable artificial intelligence (XAI) in the loss function. In a classification task, XAI generates explanations indicating which parts of the input were considered the most critical in assigning the input to a label. Moreover, it is possible to use the model's gradients to derive these explanations [14]. A widely used method is Saliency [15], which creates explanations by calculating the gradients of the output concerning the input features. It allows for highlighting of regions where a slight change in the input would significantly change the prediction. However, this approach can produce noisy explanations. The Gradient⊙Input [16] addresses this issue by calculating the elementwise multiplication of gradients by the input. The input acts as a model-independent filter, which reduces noise and smooths the explanations. Another widely used method is DeepLIFT [16], which attributes importance to the input features by comparing the activations that the actual input and a reference input cause in each neuron. It uses the difference between the activations as importance scores for the input features. Gradient-based XAI generates detailed, fine-grained explanations at the pixel level with significant computational efficiency compared to XAI based on feature perturbation [17,18].

The images generated by a combination of GAN and XAI in the loss function can be used in a data augmentation task and potentially improve the classification performance of state-of-the-art methods, such as Transformer-based models [19–22]. These methods leverage self-attention mechanisms to enable holistic image understanding and achieve top-notch performance in visual recognition tasks [23–29]. ViT [19] is a prime example. It operates on a patch-based representation of images using the self-attention mechanism to capture global dependencies and learn long-range relationships between the patches. PVT [20] introduces a hierarchical approach operating at different spatial resolutions to capture fine-grained details and high-level semantic information. It uses local–global and global–local attention modules. The local–global attends to local and global features within the same input region, while the global–local attends to the global representation while incorporating local information. DeiT [21] focuses on training efficiency on smaller datasets. It employs data augmentation and knowledge distillation, in which a teacher model guides the training of the Transformer-based student model. A learnable distillation and class tokens allow the student to learn from the original and the teacher's predictions. CoAtNet [22] is a hybrid architecture that uses convolutional layers to extract local features in the initial stages and comprises self-attention layers to model long-range dependencies and global context within the image in later stages. Positional encodings are added to the embeddings in the attention stages to retain spatial information, which helps the model understand the relative positions of features within the image.

Therefore, considering the advances previously described, we present a new way to train GANs using XAI in backpropagation. Our approach involves extracting XAI explanations from D to identify the most critical features of the input and feeding this information into G via the loss function. We used traditional architectures as a basis and modified the loss function to propagate a matrix instead of just an error value. This matrix was derived from the explanations and the discriminator error. We investigated the proposal's relevance in the histopathology context, which is known to be challenging due to the low availability of labeled images caused by privacy concerns and labeling costs [30–34]. We performed the data augmentation of relevant datasets, such as breast, colon, and liver histological images, and classified them using Transformers. Through experiments, we show that our proposal improves the quality and variability of the artificial images compared to traditional GANs, promoting an increase in the generalization and classification performance of state-of-the-art Transformer-based methods. This focus is essential to understanding the added value that XAI brings to GANs, particularly in providing valuable feature information to G during the training. Our work aimed to fill a specific gap in the literature by demonstrating how XAI can be a powerful tool in enhancing GAN performance. This research makes the following significant contributions:

1. An approach that feeds G with substantial information concerning the images' features, increasing the quality and variability of the generated images.

2. A training strategy that produce images with more realistic features, promoting an increase in the generalization and classification performance of state-of-the-art Transformer-based methods.
3. The indication of the best combination between the GAN and XAI to generate and classify the histological images explored here.

2. Methodology

We named the proposed method XGAN, and a schematic summary of its structure is illustrated in Figure 1. It comprises a generator (G) and a discriminator (D). G receives a random signal vector z and outputs an image $G(z)$, while D classifies authentic x and artificial images $G(z)$. The model uses XAI to extract feature information from D and feed it back to G to perform a new form of training called educational training. To conduct this training, we propose a new loss function $\mathcal{L}_G^{ed}$ that uses traditional adversary losses ($\mathcal{L}_G^{adv}$) combined with XAI explanations (E) to backpropagate important information to the generator. The new loss function was defined as follows:

$$\mathcal{L}_G^{ed} = \mathcal{L}_G^{adv} * E, \tag{1}$$

in which $*$ is the multiplication operation.

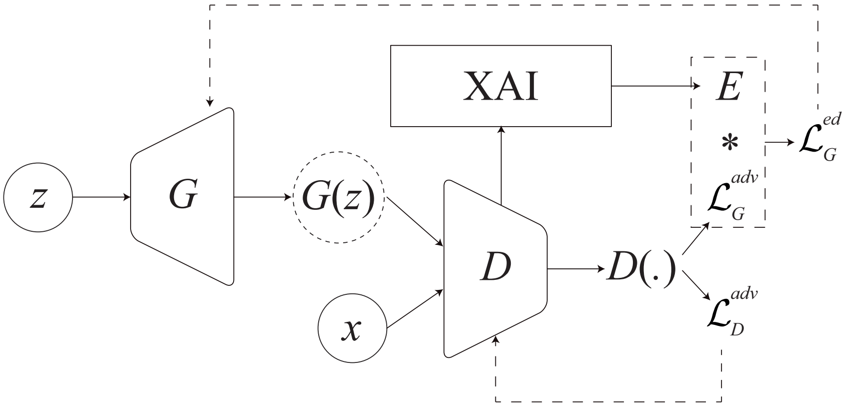

Figure 1. Schematic summary of the proposed model.

The gradient determines how much to adjust each weight of G so that the loss function walks towards the optimum. Incorporating E within the gradient enables emphasis on areas corresponding to objects of interest while dampening the influence of less relevant regions. Our method proposes a student-versus-teacher relationship. In this relationship, E corresponds to a test answer in which the professor (D) informs the student (G) of their test score, indicating features drawn close to reality and those not similar to the original images. Thus, instead of propagating just one value that indicates D's classification error, we propagate a matrix with relevant information for each pixel in the image.

To propagate a matrix, an operation known as the vector-Jacobian product is required, defined as

$$J \cdot \vec{v}, \tag{2}$$

where $\vec{v}$ is a multidimensional vector of the same dimension as the explanations E with 1 in all positions, and J is the Jacobian matrix, a matrix of partial derivatives that indicates how the output changes concerning the input. AutoDiff uses the Jacobian matrix to perform the backpropagation process by stacking the partial derivatives for each output concerning each input variable. Considering that a neural network is a function $f : \mathbb{R}^n \to \mathbb{R}^m$ that maps n-dimensional input vectors (x) to m-dimensional output vectors (y), the matrix J is defined as

$$J = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}. \tag{3}$$

The Jacobian matrix indicates how the output changes when a small amount of the input changes. In the proposed method, J also informs the change that each pixel of the artificial image causes in the prediction of D, assigning greater weights to the more relevant pixels via E.

We used the DCGAN, WGAN-GP, and RAGAN models to define the adversarial loss functions $\mathcal{L}_D^{adv}$ and $\mathcal{L}_G^{adv}$, and the XAI methods Saliency, DeepLIFT, and Gradient$\odot$Input to generate the explanations E. We give more details about the models and methods in the following sections.

2.1. Adversarial Loss Functions

2.1.1. DCGAN

We calculate the DCGAN adversarial loss for D ($\mathcal{L}_D^{\text{DCGAN}}$) through the binary cross-entropy:

$$\mathcal{L}_D^{\text{DCGAN}} = \mathbb{E}_{x \sim p(x)}[\log(D(x))] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))], \tag{4}$$

where $D(x)$ is D's output for real samples x, z is a random noise vector, and $D(G(z))$ is D's output for the generated images $G(z)$. For real samples x, D tries to maximize the probability of assigning them a value close to 1, while for artificial samples $G(z)$, D tries to assign a value close to zero. In contrast, the $\mathcal{L}_G^{\text{DCGAN}}$ tries to maximize the probability of D assigning a value close to 1 to the generated samples; it was defined as

$$\mathcal{L}_G^{\text{DCGAN}} = \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]. \tag{5}$$

2.1.2. WGAN-GP

The $\mathcal{L}_D^{\text{WGAN-GP}}$ was defined in terms of the Wasserstein distance $\mathcal{L}_W$ and the gradient penalty $\mathcal{L}_{GP}$:

$$\mathcal{L}_D^{\text{WGAN-GP}} = -\mathcal{L}_W + \mathcal{L}_{GP}, \tag{6}$$

in which $\mathcal{L}_W$ is the difference between the expected values of D's output for real and generated samples:

$$\mathcal{L}_W = \mathbb{E}_{x \sim p(x)}[D(x)] - \mathbb{E}_{z \sim p(z)}[D(G(z))], \tag{7}$$

and

$$\mathcal{L}_{GP} = \lambda \mathbb{E}_{\hat{x} \sim p(\hat{x})}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{8}$$

where $\hat{x}$ is a sample along a straight line between a real sample and a generated sample, and λ is a hyperparameter that controls the strength of the penalty.

For G, $\mathcal{L}_G^{\text{WGAN-GP}}$ was defined as the negation of the expected value of D's output for generated samples:

$$\mathcal{L}_G^{\text{WGAN-GP}} = -\mathbb{E}_{z \sim p(z)}[D(G(z))]. \tag{9}$$

2.1.3. RAGAN

The $\mathcal{L}_D^{\text{RAGAN}}$ was defined as the sum of the DCGAN loss and the relativistic discriminator loss:

$$\mathcal{L}_D^{\text{RAGAN}} = \mathcal{L}_D^{\text{DCGAN}} + \mathcal{L}_{rel}, \tag{10}$$

where

$$\mathcal{L}_{rel} = -\frac{1}{2}\mathbb{E}_{x\sim p(x),z\sim p(z)}[\log(D(x) - D(G(z))]. \tag{11}$$

The $\mathcal{L}_G^{\text{RAGAN}}$ was defined as

$$\mathcal{L}_G^{\text{RAGAN}} = -\frac{1}{2}\mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z))] - \mathbb{E}_{x\sim p(x)}[\log(D(x)]. \tag{12}$$

2.2. XAI Methods

For this work, we used gradient-based XAI techniques to extract the most critical features from D's gradients. We opted to use this type of XAI due to its computational efficiency and capacity for creating fine-grained pixel-level explanations [14]. We used the Saliency, DeepLIFT, and Gradient⊙Input methods to generate the explanations E.

To calculate the Saliency [15] explanation (E_{Saliency}) for a fake image $G(z)$, we calculated the partial derivative of associated output $D(G(z))$ concerning the input $G(z)$:

$$E_{\text{Saliency}} = \frac{\partial D(G(z))}{\partial G(z)}. \tag{13}$$

To determine the Gradient⊙Input [16] explanation ($E_{\text{Gradient}\odot\text{Input}}$), we calculated the elementwise multiplication of gradients by the input:

$$E_{\text{Gradient}\odot\text{Input}} = \frac{\partial D(G(z))}{\partial G(z)} \odot G(z). \tag{14}$$

Finally, to define DeepLIFT [16], we calculated the importance scores of the input ($G(z)$) by comparing their contributions to the output against a reference input (x^0). We used the minimal activation, that is, all zeros, as a reference. Thus, considering t as an output neuron and $\eta_1, \eta_2, \cdots, \eta_n$ as the set of neurons necessary to calculate t, $\Delta t = t - t^0$ is the difference between the outputs caused by $G(z)$ and x^0. We calculated the explanation E_{DeepLIFT} as follows:

$$E_{\text{DeepLIFT}} = \sum_{i=1}^{n} C_{\Delta\eta_i \Delta t} = \Delta t, \tag{15}$$

where $\Delta\eta_i$ is the difference between neuron activations caused by $G(z)$ and x^0, and $C_{\Delta\eta_i\Delta t}$ is the contribution score of $\Delta\eta_i$ to Δt.

2.3. Datasets

The CR dataset [35] (Figure 2) consists of 165 RGB images of colorectal tissue obtained from 16 representative sections of colorectal cancer at stages T3 or T4. The samples are divided between benign (74 images) and malignant tumors (91 images). Image acquisition was performed by digitally photographing histological sections with a Zeiss MIRAX MIDI slide scanner. The pixel resolution was 0.620 µm, corresponding to a 20× magnification. The images have different sizes, ranging from 567 × 430 to 775 × 522 pixels.

The LA and LG datasets [36] (Figures 3 and 4) comprise RGB images of liver tissue obtained from mice. The LG consists of 265 images obtained from male (150) and female (115) mice subjected to calorie-restricted diets. The LA dataset is composed of 529 images divided into four classes, each representing a different age group of female mice on ad libitum diets: 1 month (100), 6 months (115), 16 months (162), and 24 months (152) of age. The samples were obtained using a Carl Zeiss Axiovert 200 microscope and a 40× objective. All images have a resolution of 417 × 312 pixels. Both datasets were available through the Atlas of Gene Expression in Mouse Aging Project (AGEMAP).

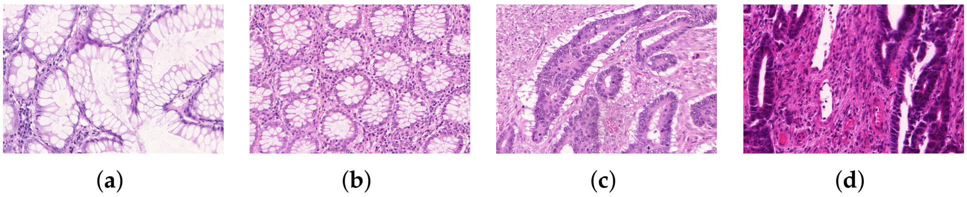

Figure 2. Example images from the CR dataset: (**a**,**b**) benign tumors, (**c**,**d**) malignant tumors.

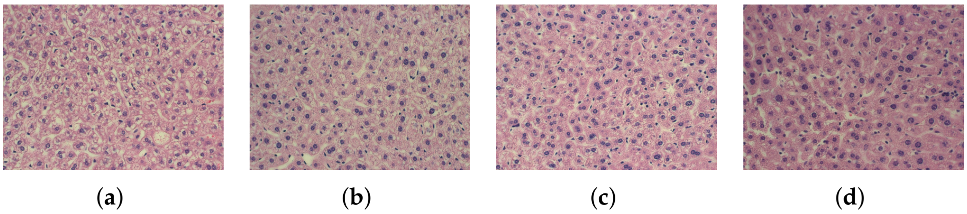

Figure 3. Example images from the LA dataset: (**a**) 1 month, (**b**) 6 months, (**c**) 16 months, and (**d**) 24 months.

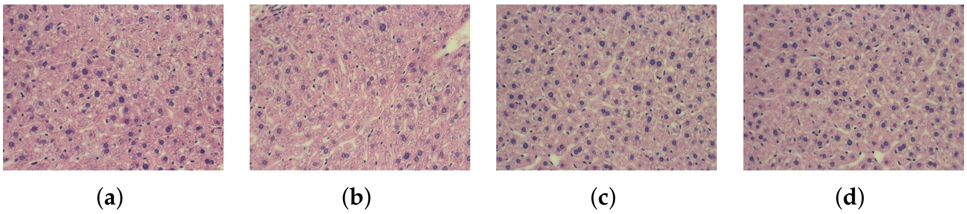

Figure 4. Example images from the LG dataset: (**a**,**b**) male and (**c**,**d**) female.

The UCSB dataset [37] (Figure 5) is a critical case of image scarcity. It consists of 58 RGB images of breast tissue divided into two groups: benign breast cancer (32) and malignant breast cancer (26). The samples provided by the Center of Bio-Image Informatics at the University of California at Santa Barbara have a quantization rate of 24 bits and a size of 768 × 896 pixels.

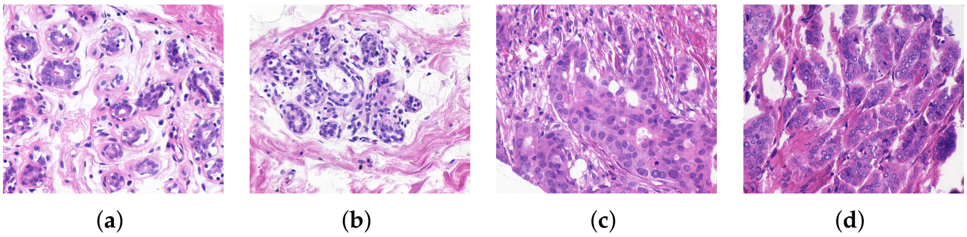

Figure 5. Example images from the UCSB dataset: (**a**,**b**) benign tumors, (**c**,**d**) malignant tumors.

2.4. Performance Evaluation

We evaluated the performance of the proposed model in two steps. First, we quantitatively assessed the quality of the artificial images using the Fréchet inception distance (FID) and the inception score (IS) metrics. Second, we performed the data augmentation and classified the images using the Transformer models ViT, PVT, DeiT, and CoAtNet for each dataset, evaluating the accuracy of each case. The idea was to compare how the XAI models can improve the quality of the generated images compared to the original architectures and how this impacts the classification of the Transformer models.

2.4.1. Image Quality Evaluation

Many metrics are available to evaluate the quality of artificial images, and each has strengths. However, interpreting the results can be challenging and requires careful consideration. For example, a specific FID variation (Gromov–Fréchet Distance) was designed to compare metric space shapes, particularly in shape analysis and topology. It has valuable applications such as shape matching or comparing complex geometric objects. However, GAN-generated images are typically evaluated based on the similarity of their distributions to authentic images, where the focus is on pixel values, textures, and high-level features captured by neural networks rather than the geometric structure of the underlying data.

Some assessments could also be performed via precision–recall (PR) metrics. These metrics are derived from the F1 score or area under the PR curve and can provide insights into a GAN's performance in generating realistic and diverse images. For instance, PR metrics can provide a valuable perspective by assessing the trade-off between the precision (how many of the generated images are relevant or high-quality) and recall (how many of the relevant images are generated) of the model. However, some disadvantages compared to FID and IS are observed. PR metrics require the definition of thresholds to determine what constitutes a realistic or diverse image, which can be complex and context-dependent. The choice of threshold can significantly affect the results, making PR metrics less consistent across different models and datasets. The interpretation of PR curves might be more complex than single-valued metrics like FID or IS.

The FID and IS are particularly pertinent for assessing image quality because they have been widely adopted in the field and are well-suited for comparing generative models by evaluating the fidelity and diversity of the generated images. FID, for instance, compares the distribution of generated images with real images, capturing the similarity in a way that aligns with human perception. IS measures how distinct and meaningful the generated images are using the classification confidence of a pre-trained network.

In this work, we applied the FID metric [38] to assess the quality of artificial images quantitatively. This metric measures the distance between the distributions of real and generated images. Thus, lower FID scores indicate higher similarity between the distributions, meaning that the generated images closely resemble the original ones. The FID measures the similarity between two multivariate Gaussian distributions, defined by the mean and covariance matrix of activation features extracted from Inception v3's 2048th layer. Mathematically, the FID score is defined by

$$FID = \|\mu_{\mathrm{r}} - \mu_{\mathrm{f}}\|^2 + \mathrm{Tr}(\Sigma_{\mathrm{r}} + \Sigma_{\mathrm{f}} - 2(\Sigma_{\mathrm{r}} \cdot \Sigma_{\mathrm{f}})^{0.5}), \tag{16}$$

where μ_{r} and μ_{f} are the mean features of real and fake images. Σ_{r} and Σ_{f} are the covariance matrices of real and fake image features, and $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

We also applied the IS metric [39] to estimate the diversity of the generated images. A higher IS suggests greater variety in the assigned classes, although it does not necessarily indicate a high degree of realism. In the IS calculation, fake images were evaluated based on the activations of the final classification layer of a pre-trained Inception v3 model. This model assigns a probability distribution to each image over predefined classes in the ImageNet dataset. Diverse images are expected to have probabilities spread across multiple classes. The IS was calculated by taking the average entropy of all generated images and computing its exponential value:

$$IS = \exp(\mathbb{E}_x[D_{\mathrm{KL}}(p(y|x)||p(y))]), \tag{17}$$

where $p(y|x)$ is the probability of class y being assigned to the generated image x, $p(y)$ is the marginal probability of class y in the dataset, $\mathbb{E}_x$ denotes the expectation taken over all generated images, $D_{\mathrm{KL}}(p(y|x)||p(y))$ is the Kullback–Leibler divergence between $p(y|x)$ and $p(y)$, and $\exp(\cdot)$ represents the exponential function.

2.4.2. Classification Evaluation

To train and evaluate the GAN and XGAN models, we employed a strategy based on regions of interest (ROIs). Figure 6 illustrates the classification evaluation process. Initially, we reserved 20% of the dataset for testing and divided the remaining 80% into five stratified folds. Each fold's images were cropped into 64 × 64-pixel ROIs to ensure that ROIs from the same image did not appear in both the training and validation sets. We conducted classification using cross-validation, with four folds for training and one fold for validation. For data augmentation, we used the GAN and XGAN models exclusively in the training group to prevent overfitting, with an augmentation rate equal to 100% of the training set size. We fine-tuned Transformer models, initially trained on the ImageNet dataset, for the datasets under investigation. Classification performance was evaluated using the accuracy metric, which measures the proportion of correctly classified samples. Finally, we compared the classification performance of Transformer models with and without data augmentation using the proposed XGAN and the original GAN architectures.

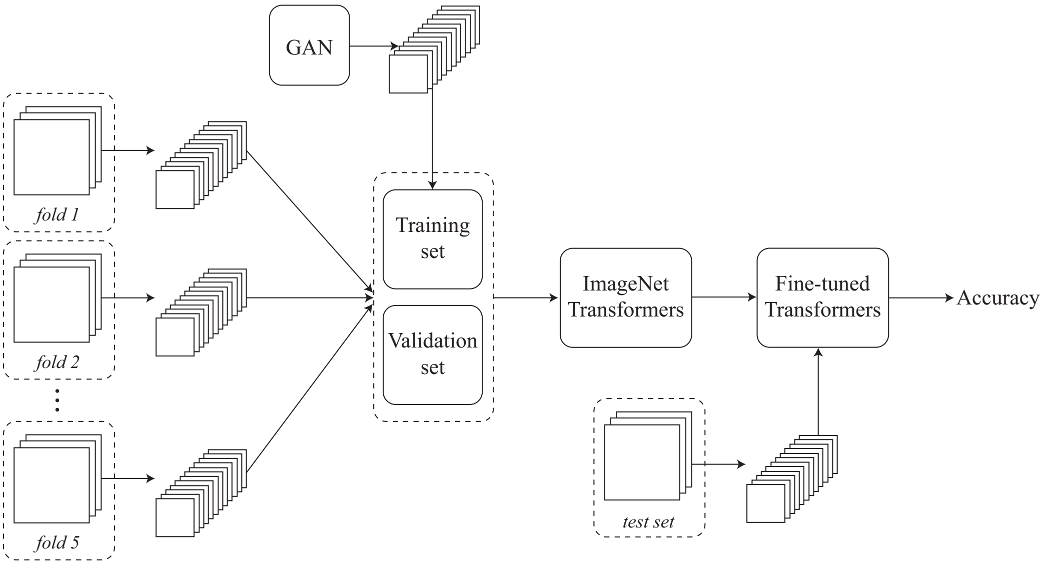

Figure 6. Schematic illustration of the classification evaluation process.

2.5. Execution Environment

The proposed method was implemented using Python 3.9.16 and the Pytorch 1.13.1 API. The experiments were performed on a computer with a 12th Generation Intel® Core™i7-12700, 2.10 GHz, NVIDIA® GeForce RTX™3090 card, 64 GB of RAM, and a Windows operating system with 64-bit architecture.

It is important to consider that we have developed our code in PyTorch, which is well known for its flexibility and strong support within the MLOps ecosystem. PyTorch's integration capabilities with various MLOps tools and platforms, such as TensorBoard, MLflow, and Kubernetes, make incorporating continuous integration and deployment (CI/CD) pipelines, model versioning, and automated monitoring straightforward. This compatibility ensures that our work can be efficiently transitioned into production environments, facilitating real-time processing and decision making.

3. Results and Discussion

We used XAI explanations to improve the training of GANs and generate artificial images with higher quality and variability. Figures 7–10 show some examples of the original images and those generated by the GAN and XGAN models for the CR, LA, LG, and UCSB datasets, respectively. We conducted experiments using the FID and IS metrics to assess quantitatively the quality and variability of the artificial images. Table 1 shows the results of these experiments. The results are organized by base architecture: DCGAN, RAGAN, and WGAN-GP. Scores in bold indicate the best results regarding the base architecture. The green and red arrows indicate whether the FID and IS obtained with XGAN are better or worse than those obtained with the original architecture.

Table 1. FID and IS scores for CR, LA, LG, and UCSB datasets. Scores in bold indicate the best results regarding the base architecture, with arrows indicating whether XGAN performs better (green arrow) or worse (red arrow) than the original architecture.

		CR		**LA**		**LG**		**UCSB**	
		FID	**IS**	**FID**	**IS**	**FID**	**IS**	**FID**	**IS**
DCGAN	No XAI	59.13	2.41	**127.94**	**1.57**	108.18	1.40	72.77	2.78
	Saliency	**57.01** ↓	**2.43** ↑	128.07 ↑	1.48 ↓	114.01 ↑	1.38 ↓	70.72 ↓	2.72 ↓
	DeepLIFT	62.45 ↑	2.38 ↓	134.21 ↑	1.46 ↓	98.24 ↓	**1.42** ↑	**62.34** ↓	**2.83** ↑
	Gradient⊙Input	60.32 ↑	**2.43** ↑	133.88 ↑	1.51 ↓	**93.88** ↓	1.39 ↓	69.90 ↓	2.72 ↓
RAGAN	No XAI	68.39	2.40	107.15	1.52	95.91	**1.45**	68.42	2.62
	Saliency	50.39 ↓	2.45 ↑	124.82 ↑	1.58 ↑	**88.30** ↓	1.42 ↓	**63.21** ↓	**2.79** ↑
	DeepLIFT	**46.02** ↓	2.50 ↑	**104.08** ↓	1.50 ↓	101.78 ↑	1.40 ↓	69.73 ↑	2.62 ↑
	Gradient⊙Input	58.57 ↓	**2.56** ↑	105.83 ↓	**1.61** ↑	95.12 ↓	1.40 ↓	66.68 ↓	2.65 ↑
WGAN-GP	No XAI	82.81	2.21	191.59	1.51	137.46	1.45	**81.07**	2.44
	Saliency	**72.01** ↓	2.28 ↑	**152.19** ↓	**1.59** ↑	**128.26** ↓	1,43 ↓	92.27 ↑	2.50 ↑
	DeepLIFT	76.86 ↓	**2.29** ↑	158.51 ↓	1.56 ↑	153.10 ↑	1,57 ↑	89.34 ↑	2.55 ↑
	Gradient⊙Input	74.50 ↓	2.17 ↓	178.23 ↓	1.35 ↓	143.97 ↑	1,50 ↑	87.65 ↑	**2.74** ↑

Considering the CR dataset (Figure 7), it is possible to note in Table 1 that XDCGAN + Saliency achieved the best FID and IS regarding the DCGAN-based architectures. The values were 57.01 and 2.43, respectively. Considering the RAGAN-based models, it is worth noting that all combinations of XRAGAN improved the FID and IS compared with RAGAN. The XRAGAN + DeepLIFT was the highlight, providing the lowest FID, 46.02, representing a 32.70% decrease compared to RAGAN (68.39). The XWGAN-GP also improved the FID in all cases. The combination XWGAN-GP + Saliency provided the lowest FID, 72.01, 13.04% lower than WGAN-GP (82.81).

On the LA dataset (Figure 8), XDCGAN did not improve the FID and IS. DCGAN achieved the best quality, with an FID of 127.94 and an IS of 1.57. However, considering RAGAN-based models, XRAGAN + DeepLIFT and XRAGAN + Gradient⊙Input improved the FID and IS slightly. The best combination was XRAGAN + DeepLIFT, which provided an FID improvement of 2.8%, 104.08 against 107.15 with RAGAN. On the other hand, XWGAN-GP improved the FID in all cases. XWGAN-GP + Saliency provided the best FID, 152.19, representing a 20.56% decrease compared to WGAN-GP.

Regarding the LG dataset (Figure 9) and the performance using DCGAN-based architectures (Table 1), XDCGAN + Gradient⊙Input achieved the best FID, 93.88, representing a 13.22% decrease compared to DCGAN (108.18). Considering RAGAN-based models, XRAGAN + Saliency and XRAGAN + Gradient⊙Input improved the FID and IS. XRAGAN + Saliency provided the best FID, 88.30, representing an 8.99% decrease compared to RAGAN (95.91). When using WGAN-GP as the base architecture, XWGAN-GP + Saliency provided the best FID, 128.26, 6.52% less than WGAN-GP (137.46).

Finally, considering the UCSB dataset (Figure 10), XDCGAN + DeepLIFT achieved an FID of 62.34 and an IS of 2.83. This combination was the best FID and IS among the DCGAN-based architectures, representing about a 14.33% FID improvement. Considering the RAGAN-based models, XRAGAN + Saliency and XRAGAN + Gradient⊙Input improved the FID and IS. XRAGAN + Saliency provided the best FID, 63.21, representing a 7.61% decrease compared to RAGAN (68.42). This combination also produced more diverse images. The IS score was 2.79. XWGAN-GP showed no improvement compared to WGAN-GP. The best case was WGAN-GP, with an FID of 81.07. However, XWGAN-GP provided the best IS, 2.74.

We used the generated images to augment the datasets and train the Transformer models. Our goal was to verify whether the XAI methods' feature information could improve the Transformer models' classification performance. We followed the strategy described in Section 2.4.2. Tables 2–5 show the classification results on the CR, LA, LG, and UCSB datasets, respectively. Each table's first row presents the classification accuracy without data augmentation (without DA), and results using a GAN in the data augmentation are organized by base architecture: DCGAN, RAGAN, and WGAN-GP. We considered the classification performance without data augmentation and with data augmentation using original architectures (DCGAN, RAGAN, and WGAN-GP with no XAI) as a baseline. We compared the results obtained with XGAN (XDCGAN, XRAGAN, and XWGAN-GP) with the defined baselines to determine whether the proposed method is capable of improving the classification performance of Transformer-based models via data augmentation and whether educational training using XAI is capable of enhancing the quality and classification performance of a given base GAN architecture. Thus, bold values indicate the best accuracy given a base architecture.

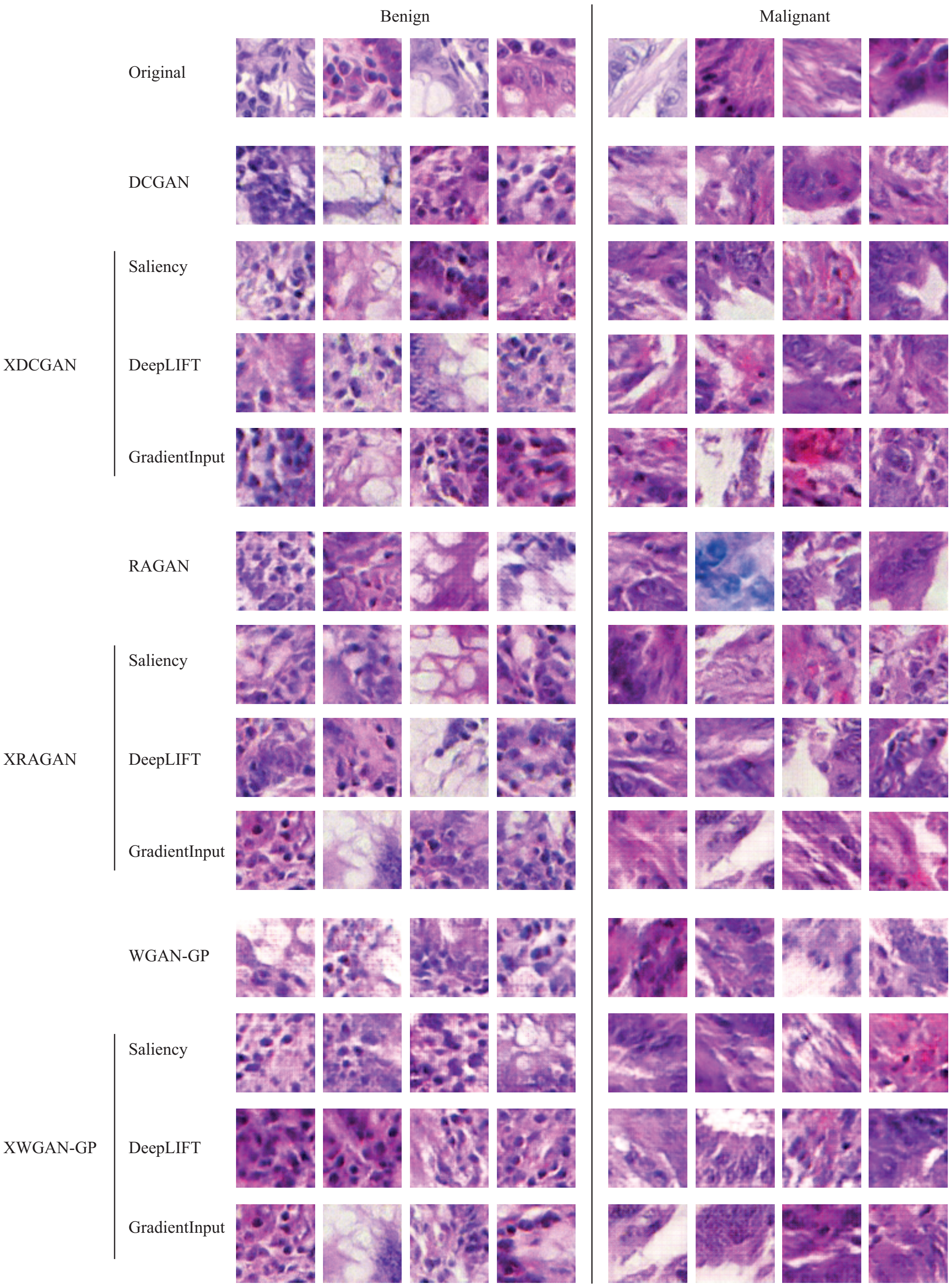

Figure 7. Examples of generated images for the CR dataset.

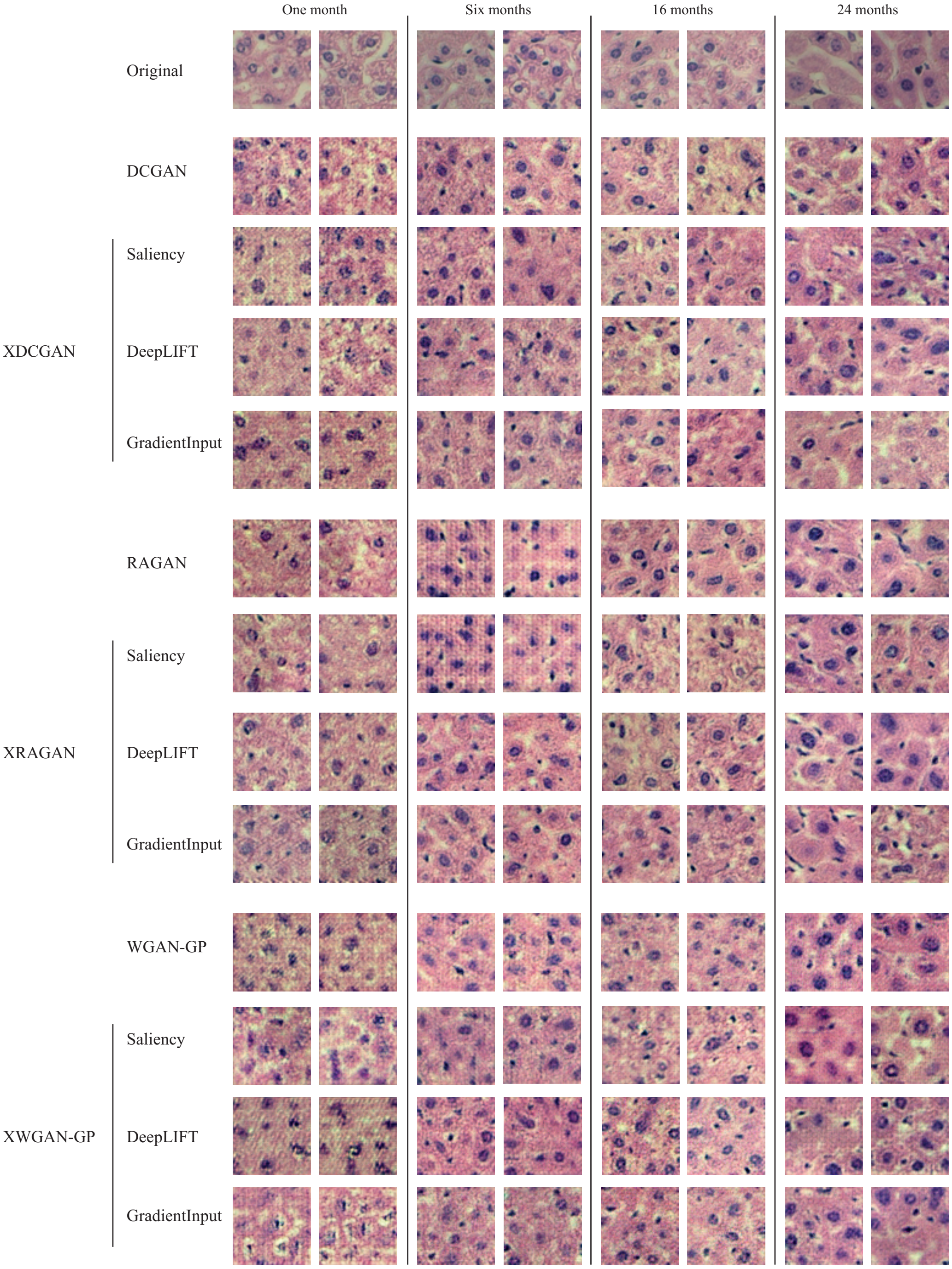

Figure 8. Examples of generated images for the LA dataset.

Considering the classification results regarding the CR dataset (Table 2), the combination with the best FID among the DCGAN-based architectures, XDCGAN + Saliency, also provided the best classification performance. This combination provided the best accuracy with all Transformer models: 84.14% with ViT, 87.70% with DEiT, 92.24% with PVT, and 93.44% with CoAtNet. With CoAtNet, XDCGAN + Gradien⊙Input also achieved an accuracy of 93.44%, the same as XDCGAN + Saliency. The IS value with this combination (2.43) may have influenced this result, promoting more significant variability in artificial images, and consequently, better model generalization. Considering the RAGAN-based models, it is possible to note that XRAGAN + DeepLIFT provided the lowest FID (46.02) and the highest accuracy in most cases, 84.30% with ViT and 91.92% with PVT. Also, XWGAN-GP + Saliency provided the lowest FID (72.01) and achieved the highest accuracy with ViT (84.41%), DEiT (88.09%), and CoAtNet (93.64%). The XWGAN-GP + Gradient⊙Input provided the second-lowest FID (74.01) and achieved the highest accuracy with PVT (92.13%).

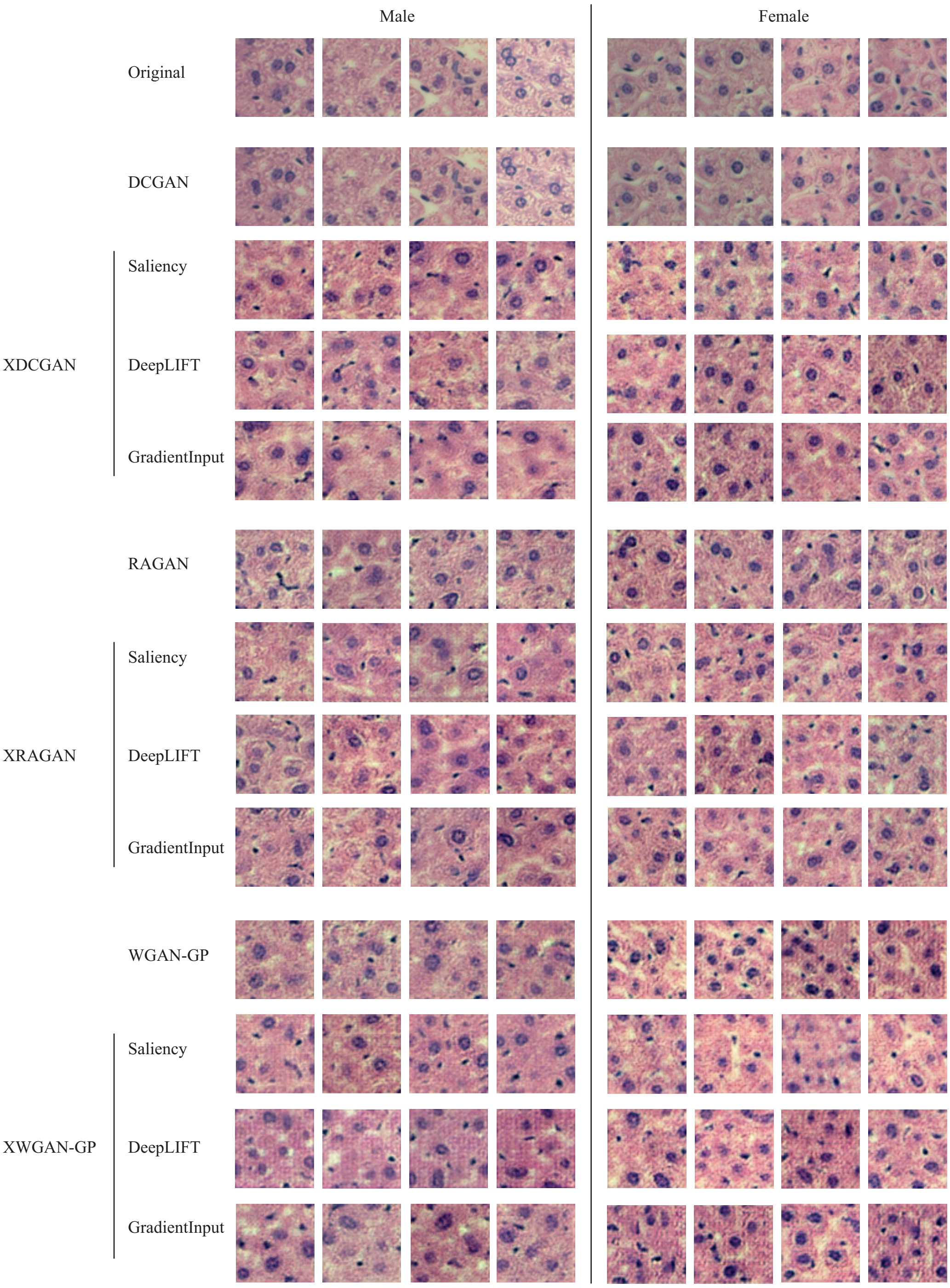

Figure 9. Examples of generated images for the LG dataset.

Regarding the LA dataset, XDCGAN could not improve the FID (Table 1). XDCGAN + Saliency, however, achieved the second-lowest FID (128.07), and it is worth noting in Table 3 that this combination provided the best accuracy with DEiT (95.45%). It is also worth noting that the architecture with no GAN improved the classification performance with ViT, which achieved 95.04% without data augmentation. Nevertheless, taking into account the RAGAN-based architectures, the combination with the best FID, the XRAGAN + DeepLIFT, provided the best classification performance with PVT, 97.71% against 97.32%

with RAGAN, and 96.44% without data augmentation. XWGAN-GP improved DEiT, PVT, and CoAtNet classification performance compared to WGAN-GP. For instance, WGAN-GP + Saliency achieved an accuracy of 97.79% with DEiT and 98.24% with PVT. This result represents an improvement of 3.81% and 1.87% compared to the models without data augmentation, respectively, and 3.01% and 1.23% compared to WGAN-GP. Also, the combination XWGAN-GP + DeepLIFT, which provided the second-best FID and IS, achieved the best accuracy with CoAtNet, 99.34%.

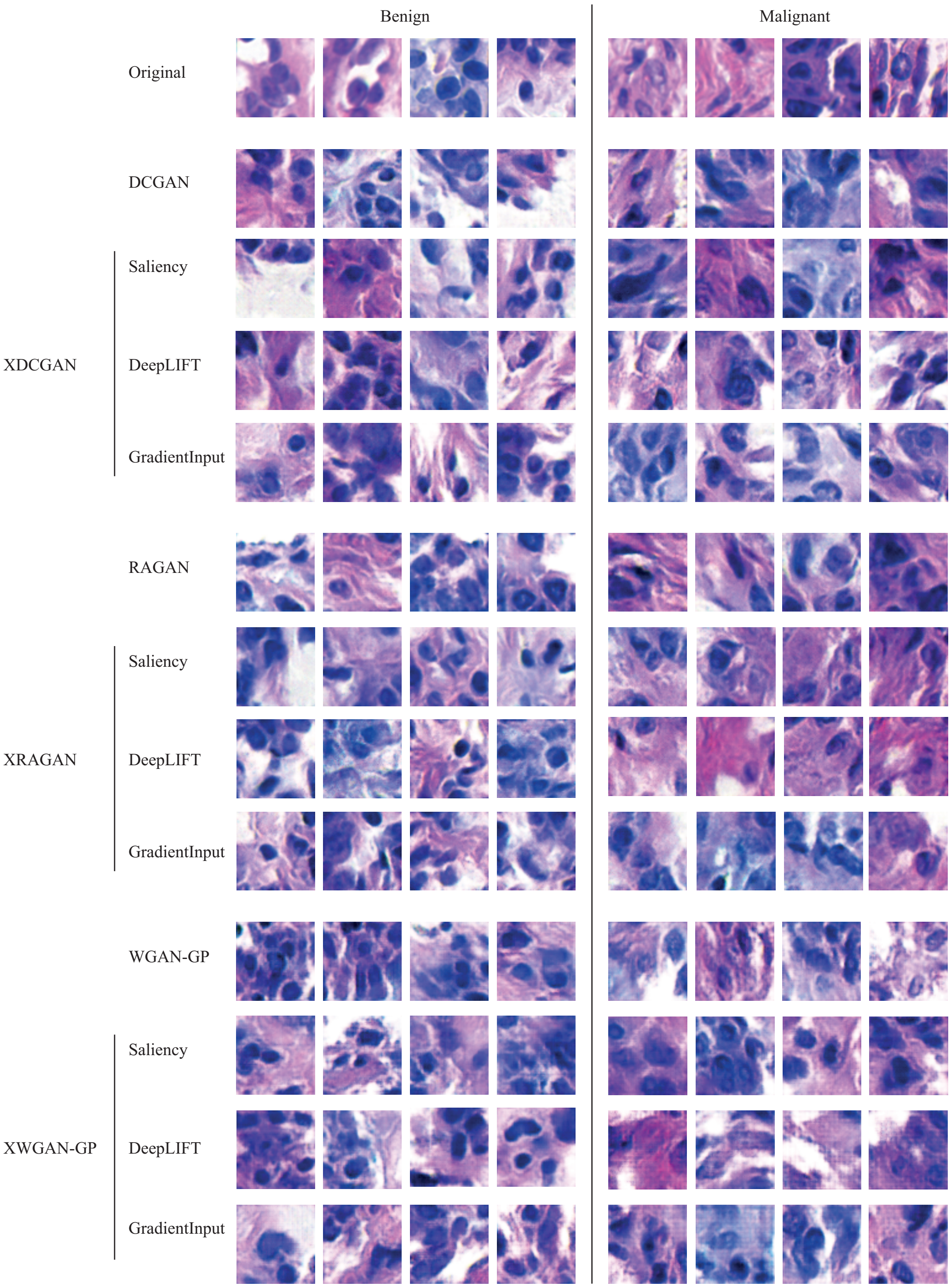

Figure 10. Examples of generated images for the UCSB dataset.

Table 2. Accuracy metric for the CR dataset.

		ViT	**DEiT**	**PVT**	**CoAtNet**
without DA		81.29%	87.89%	91.81%	93.19%
DCGAN	No XAI	82.67%	87.48%	91.80%	92.30%
	Saliency	**84.14%**	**87.70%**	**92.24%**	**93.44%**
	DeepLIFT	79.97%	87.63%	91.48%	93.15%
	Gradient⊙Input	82.89%	87.25%	91.98%	**93.44%**
RAGAN	No XAI	83.53%	**88.44%**	91.02%	92.73%
	Saliency	81.87%	88.33%	90.98%	**93.88%**
	DeepLIFT	**84.39%**	87.28%	**91.92%**	93.06%
	Gradient⊙Input	80.56%	87.28%	91.86%	92.64%
WGAN-GP	No XAI	81.60%	86.52%	91.15%	91.95%
	Saliency	**84.41%**	**88.09%**	91.98%	**93.64%**
	DeepLIFT	78.25%	87.91%	91.54%	93.36%
	Gradient⊙Input	83.34%	87.96%	**92.13%**	93.13%

Table 3. Accuracy metric for the LA dataset.

		ViT	**DEiT**	**PVT**	**CoAtNet**
without DA		95.04%	94.2%	96.44%	99.12%
DCGAN	No XAI	93.14%	94.91%	**97.75%**	**99.32%**
	Saliency	92.29%	**95.45%**	96.98%	98.93%
	DeepLIFT	93.46%	94.38%	97.16%	99.5%
	Gradient⊙Input	92.33%	93.48%	96.83%	99.26%
RAGAN	No XAI	94.74%	94.34%	97.32%	**99.33%**
	Saliency	92.61%	94.04%	97.42%	98.44%
	DeepLIFT	94.00%	93.33%	**97.71%**	99.22%
	Gradient⊙Input	93.77%	**94.98%**	97.65%	99.16%
WGAN-GP	No XAI	94.04%	94.93%	97.05%	99.07%
	Saliency	94.09%	**97.79%**	**98.24%**	98.77%
	DeepLIFT	94.36%	97.46%	97.63%	**99.34%**
	Gradient⊙Input	94.73%	95.46%	97.56%	98.49%

Table 4 shows the accuracy of the LG dataset. Once again, no architecture was able to improve classification with ViT. However, the two best XDCGAN models in terms of FID (Table 1), XDCGAN + Gradient⊙Input (99.33) and XDCGAN + DeepLIFT (98.24), achieved the highest accuracy with PVT (99.20%) and DEiT (97.43%), respectively. The XRAGAN model showed improvements in FID and IS when combined with Saliency (Table 1). This combination achieved the highest accuracy with DEiT (96.77%) and PVT (99.08%). Also, the combination XRAGAN + Gradient⊙Input achieved 99.79% with CoAtNet. The XWGAN-GP model also improved FID and IS when combined with Saliency. It achieved the highest accuracy with DEiT (97.17%) and CoAtNet (99.81%). The XWGAN-GP + Gradient⊙Input combination provided the second-lowest FID and achieved the highest accuracy with PVT (98.94%).

Table 5 shows the accuracy of the UCSB dataset. The XDCGAN + Gradient⊙Input achieved the best results with DEiT (75.14%) and PVT (78.36%). Regarding the XRAGAN models, XRAGAN + Saliency and XRAGAN + Gradient⊙Input achieved the lowest FID (Table 1). XRAGAN + Saliency achieved the highest accuracy with DEiT (75.88%) and XRAGAN + Gradient⊙Input with PVT (77.36%). The XWGAN-GP + Gradient⊙Input combination provided the lowest FID (87.65) and achieved the highest accuracy with DEiT (77.42%), PVT (77.77%), and CoAtNet (76.59%).

Table 4. Accuracy metric for the LG dataset.

		ViT	DEiT	PVT	CoAtNet
without DA		97.05%	94.88%	98.63%	99.50%
DCGAN	No XAI	96.56%	96.51%	98.80%	99.48%
	Saliency	95.64%	96.72%	99.01%	**99.86%**
	DeepLIFT	95.47%	**97.43%**	96.82%	99.34%
	Gradient⊙Input	95.78%	96.82%	**99.20%**	99.50%
RAGAN	No XAI	95.24%	97.05%	98.73%	99.74%
	Saliency	95.50%	**96.77%**	**99.08%**	99.69%
	DeepLIFT	95.33%	96.42%	98.73%	98.02%
	Gradient⊙Input	95.19%	96.53%	98.94%	**99.79%**
WGAN-GP	No XAI	91.11%	95.99%	98.66%	99.60%
	Saliency	96.32%	**97.17%**	98.87%	**99.81%**
	DeepLIFT	96.79%	96.75%	98.89%	99.17%
	Gradient⊙Input	95.19%	96.53%	**98.94%**	99.79%

Table 5. Accuracy metric for the UCSB dataset.

		ViT	DEiT	PVT	CoAtNet
without DA		75.00%	73.36%	77.27%	76.17%
DCGAN	No XAI	72.21%	74.59%	77.84%	75.01%
	Saliency	72.36%	74.78%	77.14%	74.49%
	DeepLIFT	72.04%	75.01%	76.50%	74.76%
	Gradient⊙Input	71.44%	**75.14%**	**78.36%**	74.03%
RAGAN	No XAI	71.30%	73.38%	77.00%	72.77%
	Saliency	72.91%	**75.88%**	75.82%	75.21%
	DeepLIFT	71.75%	73.08%	76.98%	74.85%
	Gradient⊙Input	71.61%	75.12%	**77.92%**	75.44%
WGAN-GP	No XAI	72.86%	75.03%	77.14%	73.54%
	Saliency	70.91%	74.54%	76.98%	72.94%
	DeepLIFT	73.51%	75.20%	76.63%	73.96%
	Gradient⊙Input	71.66%	**77.42%**	**77.77%**	**76.59%**

Based on the classification results, we analyzed the best combinations of XAI and GAN for the histological datasets. This analysis aimed to determine the number of cases in which the XGAN combinations provided the best performance given a base architecture. Table 6 shows a ranking of the best combinations. Considering all datasets, the combination XWGAN-GP + Saliency outperformed WGAN-GP in seven cases. The second and third places also included the Saliency method: XDCGAN + Saliency, with six cases; and XRAGAN + Saliency, with four cases. The subsequent three cases were combinations with the Gradient⊙Input and DeepLIFT methods: XWGAN-GP + Gradient⊙Input with five cases, XDCGAN + Gradient⊙Input with four cases, and XRAGAN + Gradient⊙Input and XRAGAN + DeepLIFT both with three cases. Finally, the two worst cases were combinations with DeepLIFT: XWGAN + DeepLIFT and XDCGAN + DeepLIFT, with only one case each. Based on these facts, the XAI method that provided the best information for the generator was Saliency, followed by Gradient⊙Input and DeepLIFT. These findings can guide researchers and experts in using GANs and XAI to develop artificial augmentation techniques.

Table 6. Ranking of the combinations, showing best cases.

Position	Combination	Number of Best Cases
1	XWGAN-GP + Saliency	7
2	XDCGAN + Saliency	6
3	XRAGAN + Saliency	4
4	XWGAN-GP + Gradient⊙Input	5
5	XDCGAN + Gradient⊙Input	4
6	XRAGAN + Gradient⊙Input	3
6	XRAGAN + DeepLIFT	3
7	XWGAN + DeepLIFT	1
7	XDCGAN + DeepLIFT	1

It is important to observe that to expand our investigation, new experiments can be performed with different models and datasets, including other types of medical images. On the other hand, incorporating more datasets from various medical fields poses significant challenges, especially when considering GAN architectures. For instance, an architecture that works well for one type of medical image may not be optimal for another, requiring extensive experimentation and customization to achieve the best results. In addition, acquiring and processing datasets from different medical domains involves overcoming various logistical, ethical, and technical hurdles, which adds to the complexity of such an expansion. In this context, our approach opens up new possibilities for enhancing data augmentation techniques and improving the overall performance of Transformer-based models in histopathological datasets. It provides new patterns and insights for specialists interested in machine learning.

Finally, we recognize the importance of discussing the computational complexity of the methods employed. However, performing a complexity analysis of algorithms, especially when integrating XAI with GANs, can be challenging due to the intricate nature of these models and the variability in computational demands across different setups. Moreover, the primary focus of our research was to explore the integration of XAI techniques with GANs to improve the quality of generated images rather than to provide a comprehensive analysis of the computational complexity. Despite the complexity analysis challenges, we consciously used gradient-based XAI methods in our research. Gradient-based XAI is the fastest among the various XAI types, making it a practical choice for our study. These methods work by calculating gradients concerning the input features, which helps to identify the input areas that most influence the model's output. This approach is computationally efficient because it leverages the gradients already computed during the backpropagation process in neural networks.

4. Conclusions

In this work, we proposed a new approach for training GANs using XAI to improve generation quality and data augmentation performance on histopathological datasets. We used XAI methods, such as Saliency, DeepLIFT, and Gradient⊙Input, to extract feature information from the discriminator and feed it to the generator during the training. We evaluated the proposed method on four histopathological datasets, CR, LA, LG, and UCSB, using the FID and IS metrics to assess the quality of generated images and the accuracy metric to compare the classification performance of four Transformer models, ViT, DEiT, PVT, and CoAtNet, with and without data augmentation. The multiple experiments provided a solid foundation to understand the effectiveness of our approach in this specific domain.

The results showed that the proposed method increased the quality and diversity of the generated images. In most cases, the XGAN provided better FID and IS values than traditional GAN models. For instance, it was possible to decrease the FID by up to 32.70% compared to the traditional architectures. This gain in quality positively affected the classification performance of the Transformer models. Accuracy was increased by up to

3.81% compared to the models without data augmentation and up to 3.01% compared to the models with traditional GAN data augmentation. We also showed that Saliency was the best method for providing information to the generator, followed by Gradient⊙Input and DeepLIFT. The XWGAN-GP + Saliency combination was a highlight, as it outperformed WGAN-GP in seven cases.

These results are significant because they show XAI's potential to improve the quality of image generation by using GANs on histopathological datasets. We demonstrated that the features provided by XAI explanations contribute to better generalization in the training of Transformer models, promoting an improvement in their classification power. Also, we identified that the saliency method provided the best features and was the most relevant method for composing a combination with GAN models. These findings provide insights and guidelines for researchers and experts interested in developing artificial augmentation techniques for histopathological datasets.

In future work, some insights can be investigated: 1. new tests using different combinations of GAN models; 2. apply other evaluation metrics, such as the precision–recall metrics, for generative models; 3. integrate MLOps approaches and pipelines for biomedical image processing; 4. tests using larger datasets with a significantly higher number of samples, including those from different medical domains, to comprehensively evaluate the performance of our proposed method and its generalization capabilities; 5. new analysis by delving deeper into the reasons behind the varying performance of different GAN and XAI combinations to provide a more thorough examination of their specific advantages and limitations; 6. statistical analysis, including significance testing, to verify the improvements in classification performance and complexity analysis of the proposed method to guide optimization processes.

Author Contributions: Conceptualization, methodology, G.B.R., A.L., M.Z.d.N. and L.A.N.; software, G.B.R., B.L.d.O.G. and V.A.T.B.; validation, formal analysis, investigation, G.B.R., A.L., B.L.d.O.G., V.A.T.B. and L.A.N.; writing—original draft preparation, G.B.R.; writing—review and editing, G.B.R., A.L., T.A.A.T., M.Z.d.N. and L.A.N.; funding acquisition, G.B.R., M.Z.d.N. and L.A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001; National Council for Scientific and Technological Development CNPq (Grants #132940/2019-1, #313643/2021-0 and #311404/2021-9); the State of Minas Gerais Research Foundation—FAPEMIG (Grant #APQ-00578-18); São Paulo Research Foundation—FAPESP (Grant #2022/03020-1).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
2. Dargan, S.; Kumar, M.; Ayyagari, M.R.; Kumar, G. A survey of deep learning and its applications: A new paradigm to machine learning. *Arch. Comput. Methods Eng.* **2020**, *27*, 1071–1092. [CrossRef]
3. Madani, A.; Moradi, M.; Karargyris, A.; Syeda-Mahmood, T. Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 1038–1042. [CrossRef]
4. Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **2018**, *321*, 321–331. [CrossRef]
5. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.

6. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
7. Nash, J.F. Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 48–49. [CrossRef]
8. Trevisan de Souza, V.L.; Marques, B.A.D.; Batagelo, H.C.; Gois, J.P. A review on Generative Adversarial Networks for image generation. *Comput. Graph.* **2023**, *114*, 13–25. [CrossRef]
9. Wang, J.; Yang, C.; Xu, Y.; Shen, Y.; Li, H.; Zhou, B. Improving gan equilibrium by raising spatial awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11285–11293.
10. Bai, Q.; Yang, C.; Xu, Y.; Liu, X.; Yang, Y.; Shen, Y. Glead: Improving gans with a generator-leading task. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 17–22 June 2023; pp. 12094–12104.
11. Wang, Z.; She, Q.; Ward, T.E. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. *ACM Comput. Surv.* **2021**, *54*. [CrossRef]
12. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
13. Jolicoeur-Martineau, A. The relativistic discriminator: A key element missing from standard GAN. In Proceedings of the International Conference on Learning Representations, Nice, France, 1–4 April 2019.
14. Nielsen, I.E.; Dera, D.; Rasool, G.; Ramachandran, R.P.; Bouaynaya, N.C. Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Process. Mag.* **2022**, *39*, 73–84. [CrossRef]
15. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014; pp. 1–8.
16. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Precup, D., Teh, Y.W., Eds.; PMLR, Proceedings of Machine Learning Research; Volume 70, pp. 3145–3153.
17. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13–17 August 2016; KDD '16, p. 1135–1144. [CrossRef]
18. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
20. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
21. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Meila, M., Zhang, T., Eds.; PMLR, Proceedings of Machine Learning Research; Volume 139, pp. 10347–10357.
22. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. In *Advances in Neural Information Processing Systems*; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 3965–3977.
23. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [CrossRef]
24. Xu, H.; Xu, Q.; Cong, F.; Kang, J.; Han, C.; Liu, Z.; Madabhushi, A.; Lu, C. Vision Transformers for Computational Histopathology. *IEEE Rev. Biomed. Eng.* **2024**, *17*, 63–79. [CrossRef]
25. Zheng, Y.; Li, J.; Shi, J.; Xie, F.; Huai, J.; Cao, M.; Jiang, Z. Kernel Attention Transformer for Histopathology Whole Slide Image Analysis and Assistant Cancer Diagnosis. *IEEE Trans. Med. Imaging* **2023**, *42*, 2726–2739. [CrossRef] [PubMed]
26. Atabansi, C.C.; Nie, J.; Liu, H.; Song, Q.; Yan, L.; Zhou, X. A survey of Transformer applications for histopathological image analysis: New developments and future directions. *Biomed. Eng. Online* **2023**, *22*, 96. [CrossRef] [PubMed]
27. Baroni, G.L.; Rasotto, L.; Roitero, K.; Tulisso, A.; Di Loreto, C.; Della Mea, V. Optimizing Vision Transformers for Histopathology: Pretraining and Normalization in Breast Cancer Classification. *J. Imaging* **2024**, *10*, 108. [CrossRef] [PubMed]
28. Goceri, E. Vision transformer based classification of gliomas from histopathological images. *Expert Syst. Appl.* **2024**, *241*, 122672. [CrossRef]
29. Mahmood, T.; Wahid, A.; Hong, J.S.; Kim, S.G.; Park, K.R. A novel convolution transformer-based network for histopathology-image classification using adaptive convolution and dynamic attention. *Eng. Appl. Artif. Intell.* **2024**, *135*, 108824. [CrossRef]
30. Xue, Y.; Ye, J.; Zhou, Q.; Long, L.R.; Antani, S.; Xue, Z.; Cornwell, C.; Zaino, R.; Cheng, K.C.; Huang, X. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. *Med. Image Anal.* **2021**, *67*, 101816. [CrossRef]

31. Inan, M.S.K.; Hossain, S.; Uddin, M.N. Data augmentation guided breast cancer diagnosis and prognosis using an integrated deep-generative framework based on breast tumor's morphological information. *Inform. Med. Unlocked* **2023**, *37*, 101171. [CrossRef]
32. Escobar Díaz Guerrero, R.; Carvalho, L.; Bocklitz, T.; Popp, J.; Oliveira, J.L. A Data Augmentation Methodology to Reduce the Class Imbalance in Histopathology Images. *J. Imaging Inform. Med.* **2024**, *37*, 1767–1782. [CrossRef]
33. Brancati, N.; Frucci, M. Improving Breast Tumor Multi-Classification from High-Resolution Histological Images with the Integration of Feature Space Data Augmentation. *Information* **2024**, *15*, 98. [CrossRef]
34. Ruiz-Casado, J.L.; Molina-Cabello, M.A.; Luque-Baena, R.M. Enhancing Histopathological Image Classification Performance through Synthetic Data Generation with Generative Adversarial Networks. *Sensors* **2024**, *24*, 3777. [CrossRef]
35. Sirinukunwattana, K.; Pluim, J.P.; Chen, H.; Qi, X.; Heng, P.A.; Guo, Y.B.; Wang, L.Y.; Matuszewski, B.J.; Bruni, E.; Sanchez, U.; et al. Gland segmentation in colon histology images: The glas challenge contest. *Med. Image Anal.* **2017**, *35*, 489–502. [CrossRef] [PubMed]
36. Shamir, L.; Orlov, N.; Mark Eckley, D.; Macura, T.J.; Goldberg, I.G. IICBU 2008: A proposed benchmark suite for biological image analysis. *Med. Biol. Eng. Comput.* **2008**, *46*, 943–947. [CrossRef] [PubMed]
37. Drelie Gelasca, E.; Byun, J.; Obara, B.; Manjunath, B. Evaluation and benchmark for biological image segmentation. In Proceedings of the 2008 15th IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 1816–1819. [CrossRef]
38. 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 *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
39. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; Chen, X. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.

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.