UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Instituto de Biociências, Letras e Ciências Exatas - Campus de São José do Rio Preto

GUILHERME BOTAZZO ROZENDO

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

São José do Rio Preto
2024

Guilherme Botazzo Rozendo

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

Tese, apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas , São José do Rio Preto, para obtenção do título de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador: Profº Dr. Leandro Alves Neves

São José do Rio Preto
2024

R893d

Rozendo, Guilherme Botazzo
Data augmentation in histopathological classification : an analysis exploring GANs with XAI-inspired models and vision transformers / Guilherme Botazzo Rozendo. -- São José do Rio Preto, 2024
71 p.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Leandro Alves Neves

1. Inteligência artificial. 2. Redes neurais (Computação). 3. Sistemas de reconhecimento de padrões. 4. Histopatologia. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

O impacto esperado do presente trabalho é o aprimoramento do treinamento de Redes Generativas Adversárias utilizadas para gerar imagens artificiais em tarefas de classificação histopatológica. Utilizamos métodos baseados em Inteligência Artificial Explicável (XAI) para extrair informações de características e incorporá-las no trainamento por meio da função de perda, uma estratégia única ainda não explorada nesse contexto. Demonstramos que essa abordagem enriquece o treinamento com informações relevantes e promove uma melhor qualidade e maior variabilidade das imagens artificiais. Na tarefa de aumento artificial de dados, essas imagens melhoram a precisão de classificação de modelos Transformer. Nosso trabalho destaca o potencial dos métodos baseados em XAI em melhorar o treinamento de GANs e sugere caminhos para futuras explorações nesta área.

POTENTIAL IMPACT OF THIS RESEARCH

The expected impact of this work is to improve the training of Generative Adversarial Networks used to generate artificial images in histopathological classification tasks. We use methods based on Explainable Artificial Intelligence (XAI) to extract feature information and incorporate it into training through the loss function, a unique strategy not yet explored in this context. We demonstrate that this approach enriches training with relevant information and promotes better quality and more significant variability of artificial images. In artificial data augmentation, these images improve the classification accuracy of Transformer models. Our work highlights the potential of XAI-based methods to improve GAN training and suggests paths for future exploration in this field.

GUILHERME BOTAZZO ROZENDO

DATA AUGMENTATION IN HISTOPATHOLOGICAL CLASSIFICATION:
An Analysis Exploring GANs with XAI-inspired models and Vision Transformers

Tese apresentado(a) à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas , São José do Rio Preto, para obtenção do título de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Data de defesa: 05/12/2024

BANCA EXAMINADORA

_____
Profº Dr. Leandro Alves Neves
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

_____
Profº Dr. Rodrigo Capobianco Guido
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

_____
Profº Dr. Fabricio Aparecido Breve
UNESP – Instituto de Geociências e Ciências Exatas de Rio Claro

_____
Profº Dr. Bruno Augusto Nassif Travençolo
Universidade Federal de Uberlândia

_____
Profº Dr. Sylvio Barbon Junior
Università degli Studi di Trieste

Dedico este trabalho à Deus.

AGRADECIMENTOS

Agradeço a Deus por me dar força e sabedoria para superar os desafios e concluir este trabalho.

Agradeço aos meus pais, familiares e amigos pelo apoio, incentivo e compreensão durante toda a minha trajetória acadêmica.

Agradeço ao meu orientador, Prof. Dr. Leandro Alves Neves, pela orientação, paciência e apoio durante todo o desenvolvimento deste trabalho.

Agradeço a Profª Drª. Alessandra Lumini pelo o apoio dedicado a este trabalho e por ter me recebido como membro de seu grupo de pesquisa na *Università di Bologna*.

Agradeço aos membros da banca examinadora, Prof. Dr. Rodrigo Capobianco Guido, Prof. Dr. Fabricio Aparecido Breve, Prof. Dr. Bruno Augusto Nassif Travençolo e Prof. Dr. Sylvio Barbon Junior, pela disponibilidade em avaliar este trabalho.

Agradeço à Universidade Estadual Paulista (UNESP) e ao Instituto de Biociências, Letras e Ciências Exatas (IBILCE) pelo apoio e estrutura oferecida durante o desenvolvimento deste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

Redes Generativas Adversárias (GANs) criam imagens colocando um gerador (G) contra um discriminador (D), visando encontrar um equilíbrio entre eles. No entanto, alcançar esse equilíbrio é desafiador porque G é treinado com base em apenas um valor que representa a previsão de D, e apenas D pode acessar as características das imagens. Introduzimos uma nova abordagem para treinar GANs usando métodos baseados em Inteligência Artificial Explicável (XAI) para melhorar a qualidade e a diversidade das imagens geradas em conjuntos de imagens histopatológicas. Utilizamos os métodos baseados em XAI para extrair informações de características de D e incorporá-las em G por meio da função de perda, uma estratégia única ainda não explorada nesse contexto. Demonstramos que essa abordagem enriquece o treinamento com informações relevantes e promove uma melhor qualidade e maior variabilidade das imagens artificiais, diminuindo até 32,7% no *Fréchet inception distance* em comparação com métodos tradicionais. Na tarefa de aumento artificial de dados, essas imagens melhoram a precisão de classificação de modelos *Transformer* em até 3,81% em comparação com modelos sem aumento artificial de dados e até 3,01% em comparação com o aumento artificial de dados tradicional de GAN. Mostramos que o método de *Saliency* fornece a G as informações de características mais informativas. O modelo permitiu destacar o potencial de métodos inspirados em XAI para melhorar o treinamento de GANs com indicações relevantes para futuras explorações nesta área.

Palavras-Chave: Redes Generativas Adversárias; Inteligência Artificial Explicável; Treinamento de GAN; Aumento de Dados; Classificação Histopatológica; Vision Transformers.

ABSTRACT

Generative Adversarial Networks (GANs) create images by pitting a generator (G) against a discriminator (D), aiming to find a balance between them. 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 methods based on Explainable Artificial Intelligence (XAI) to enhance the quality and diversity of generated images in histopathological datasets. We leverage XAI-based methods 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 of the artificial images, decreasing up to 32.7% in Fréchet inception distance 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. Our finding shows that the Saliency method provides G with the most informative feature information. This model allowed highlighting the potential of XAI-inspired methods to improve the training of GANs with relevant indications for future explorations in this area.

Keywords: Generative Adversarial Networks; Explainable Artificial Intelligence; GAN Training; Data Augmentation; Histopathological Classification; Vision Transformers.

LIST OF FIGURES

Figure 1 Examples of a histopathological images of colon (a), breast (b), and liver tissue samples stained with Hematoxylin and Eosin. 20
Figure 2 Schematic example of a spatial filter. 22
Figure 3 Example of max pooling operation with stride = 2. 22
Figure 4 Simplified illustration of the biological neuron (a) and the artificial neuron (b). 23
Figure 5 Illustrative diagram of connections between two FC layers. 24
Figure 6 Illustrative diagram of connections between two FC layers. 25
Figure 7 Schematic illustration of the ViT architecture. 27
Figure 8 Schematic illustration of a CAD system. 27
Figure 9 Schematic illustration of the *vanilla*-GAN. 28
Figure 10 Schematic summary of the proposed model. 39
Figure 11 Examples of samples from the MNIST dataset. 42
Figure 12 Examples of samples from the CIFAR-10 dataset of the classes airplane, automobile, and cat. 43
Figure 13 Example images from the CR dataset: (a) and (b) benign tumors, (c) and (d) malignant tumors. 44
Figure 14 Example images from the LA dataset: (a) one month, (b) six months, (c) 16 months, and (d) 24 months. 45
Figure 15 Example images from the LG dataset: (a) male and (b) female. 46
Figure 16 Example images from the UCSB dataset: (a) and (b) benign tumors, (c) and (d) malignant tumors. 47
Figure 17 Schematic illustration of the classification evaluation process. 47
Figure 18 Examples of artificial images generated by WGAN-GP (a), and by XWGAN-GP with the Saliency (b), DeepLIFT (c) and Gradient$\odot$Input (d) methods. 51
Figure 19 Examples of generated images for the CR dataset. 53
Figure 20 Examples of generated images for the LA dataset. 54
Figure 21 Examples of generated images for the LG dataset. 55
Figure 22 Examples of generated images for the UCSB dataset. 56

LIST OF TABLES

Table 1 – FID and IS scores for MNIST and CIFAR10 datasets. 50
Table 2 – FID and IS scores for CR, LA, LG, and UCSB datasets. 52
Table 3 – Accuracy metric for the CR dataset. 57
Table 4 – Accuracy metric for the LA dataset. 58
Table 5 – Accuracy metric for the LG dataset. 59
Table 6 – Accuracy metric for the UCSB dataset. 59
Table 7 – Rank of the combinations with best cases. 60

LIST OF ABBREVIATIONS AND ACRONYMS

CIFAR-10	*Canadian Institute For Advanced Research*
CNN	*Convolutional Neural Network*
CycleGAN	*Cycle-Consistent Generative Adversarial Network*
DCGAN	*Deep Convolutional Generative Adversarial Network*
DeepLIFT	*Deep Learning Important Features*
FC	*Fully Connected*
GAN	*Generative Adversarial Network*
Grad-CAM	*Gradient-weighted Class Activation Mapping*
LIME	*Local Interpretable Model-agnostic Explanations*
MNIST	*Modified National Institute of Standards and Technology*
RAGAN	*Relativistic Average Generative Adversarial Network*
ReLU	*Rectified Linear Unit*
WGAN	*Wasserstein Generative Adversarial Network*
XAI	*Explainable Artificial Intelligence*

LIST OF SYMBOLS

$\mathcal{L}$	Loss Function
$\mathcal{L}_{GAN}$	Adversarial Loss
$\mathcal{L}_{XAI}$	XAI Loss
$\odot$	Hadamard Product
D	Discriminator
G	Generator
J	Jacobian Matrix

CONTENTS

1	**INTRODUCTION**	**15**
1.1	Motivations	17
1.2	Objectives	19
1.3	Contributions	19
2	**THEORETICAL FOUNDATION**	**20**
2.1	Histology Images	20
2.2	Convolutional Neural Networks	21
2.2.1	**Pattern Recognition**	**21**
2.2.2	**Classification**	**23**
2.2.3	**Training a CNN**	**24**
2.3	Vision Transformers	26
2.4	Computer-aided Diagnosis	26
2.5	Generative Adversarial Networks	28
2.6	AutoDiff	29
2.7	Data Augmentation	30
2.8	Explainable Artificial Intelligence	30
2.9	Evaluation of Artificial Image Quality	31
2.10	Classification Performance Evaluation	32
3	**RELATED WORK**	**33**
3.1	Generative Adversarial Networks	33
3.2	Explanable Artificial Intelligence	35
3.2.1	**Methods that Explain a Classification Decision**	**35**
3.2.2	**Methods that Interpret Intermediate Layers**	**36**
3.3	Transformer-based Models	36
3.4	Limitations and Challenges	37
4	**METHODOLOGY**	**39**
4.1	Method Overview	39
4.2	Adversarial Loss Functions	40
4.2.1	**DCGAN**	**40**
4.2.2	**WGAN-GP**	**40**
4.2.3	**RAGAN**	**41**
4.3	XAI-based Methods	41
4.4	Datasets	42

4.4.1 MNIST and CIFAR-10 . . . 42
4.4.2 Histopathological Images . . . 43
4.4.2.1 Colorectal Cancer . . . 43
4.4.2.2 Liver Tissue . . . 43
4.4.2.3 Breast Cancer . . . 44
4.5 Image Quality Evaluation . . . 44
4.6 Classification Evaluation . . . 46
4.7 Execution Environment . . . 46

5 RESULTS AND DISCUSSION . . . 49
5.1 Results on MNIST and CIFAR10 . . . 49
5.2 Results on Histopathological datasets . . . 51

6 CONCLUSIONS . . . 61
6.1 Research Contributions . . . 61

BIBLIOGRAPHY . . . 64

1 INTRODUCTION

The effectiveness of machine learning systems relies heavily on how the analyzed data is represented. In computer vision, data has traditionally been represented using handcrafted methods specifically designed to extract features from images (Humeau-Heurtier, 2019; Salau; Jain, 2019). However, this approach often demands significant effort from specialists to create and implement methods that only sometimes deliver the expected level of performance (Bengio; Courville; Vincent, 2013; LeCun; Bengio; Hinton, 2015; Zare; Alebiosu; Lee, 2018).

In recent years, with the increase in computational power of hardware devices, representation learning algorithms have emerged as a relevant alternative to handcrafted methods. Representation learning methods are advantageous because they automatically transform raw data, such as digital image pixels, into a feature vector. These methods allow the best representation for the data to be learned automatically through a training process. In this context, deep learning-based models stand out among the different representation learning approaches, as they involve multiple processing layers and, thus, learn representations with multiple levels of abstraction (LeCun; Bengio; Hinton, 2015; Dargan et al., 2020; Anwar et al., 2018; Wang et al., 2021a).

Among the different types of deep learning techniques, artificial neural network methods stand out, where the processing layers are mathematical models that simulate a network of neurons in the human brain. The connections between neurons are defined by weights, so during the network training phase, the most relevant connections (weights) are reinforced to obtain the best data representation (Jain; Mao; Mohiuddin, 1996). The adjustment of weights is performed through an algorithm called backpropagation, which allows the error to propagate through all layers to minimize it in the next iteration. However, the proper training of artificial neural networks requires many labeled data. This issue makes the application of these methods challenging, as some real-world application contexts suffer from the low availability of labeled samples and adverse conditions, such as class imbalance, reduced number of samples, and lack of standardization in dimensions and formats. These factors can compromise the proper training of neural networks and cause overfitting (Madani et al., 2018; Frid-Adar et al., 2018). These problems and the representation power provided by artificial neural networks motivated the emergence of automatic data generation techniques, such as generative adversarial networks (GAN) (Goodfellow et al., 2014; Motamed; Rogalla; Khalvati, 2021; Ding; Huang; Cui, 2024).

GANs are unsupervised models consisting of two neural networks: a generator (G) and a discriminator (D). The purpose of G is to learn the probability distribution function of a set of examples and generate new images based on this function. On the other hand, D receives real and generated images as input and aims to distinguish between them. Both networks are trained simultaneously with different objectives. G is trained to produce realistic images to deceive D and maximize classification errors. In contrast, D is trained to identify authentic and artificial images better, minimizing classification errors. This adversarial training strategy is rooted in

game theory. It aims to achieve a balance known as the Nash equilibrium (Nash, 1950), where G generates images indistinguishable from real ones, and D can no longer differentiate between them (Creswell et al., 2018). The weight updates of G are based on D's classification error through a process called gradient descent facilitated by the backpropagation algorithm.

The advantage of GANs is their ability to synthesize high-fidelity images without explicitly defining the probability density function. Additionally, the adversarial training between the generator and discriminator models allows unlabeled images to be included in the training process, overcoming the limitation of supervised learning methods that require a considerable amount of labeled images for training (Yi; Walia; Babyn, 2019; Jabbar; Li; Omar, 2021). In computer vision, representations learned through GANs have significantly contributed to various applications, such as data augmentation (Lorencin et al., 2021; Motamed; Rogalla; Khalvati, 2021; Liao et al., 2022; Ding; Huang; Cui, 2024), classification (Apostolopoulos; Papathanasiou; Panayiotakis, 2021; Ahmad et al., 2022), and segmentation (Sun et al., 2022; Çelik; Talu, 2022). However, training GANs is challenging due to issues related to backpropagation, specifically in how the weights of G are updated. During the backpropagation process, the gradients of G are calculated from D's predictions, and G has no information on how the images' features contribute to the classification. This issue makes adversarial training an unbalanced game where D has the advantage over G (Wang et al., 2022; Bai et al., 2023). Consequently, D tends to assign higher scores to original images during training, and G fails to fool D even after the model converges (Wang et al., 2022; Souza et al., 2023).

Studies have traditionally focused on enhancing GAN training by adjusting the discriminator in the literature. For example, in DCGAN, the first GAN proposal with convolutional layers, the authors employed batch normalization and leaky ReLU activations between the intermediate layers of the discriminator to improve its stability (Wang; She; Ward, 2021). However, using the Jensen-Shannon divergence (Goodfellow et al., 2014) as the loss function in DCGAN can lead to mode collapse and vanishing gradients (Wang; She; Ward, 2021; Souza et al., 2023). WGAN-GP (Gulrajani et al., 2017) utilizes the Wasserstein distance as a loss function to address these issues, providing a more meaningful measure of the difference between probability distributions. WGAN-GP also introduces a gradient penalty term that penalizes the norm of the discriminator's gradients, thereby promoting more diverse generated samples by enforcing a more uniform distribution of gradients throughout the data space. RAGAN (Jolicoeur-Martineau, 2019) introduces the relativistic discriminator, which estimates the probability that an authentic sample is more realistic than a fake sample and vice versa. This approach considers the relative realness of real and fake samples, offering more informative feedback to the generator. RAGAN provides a more nuanced signal to the generator, enabling it to understand better how to autonomously generate plausible samples that align with the actual data distribution.

The latest techniques focus on enhancing the generator's training rather than the discriminator. These approaches involve providing more information about the features of the images to G and making the competition between the generator and discriminator more balanced. For example,

Wang et al. (2022) suggested a training technique that enhances the spatial awareness of G by 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 to align the spatial awareness of G with D's attention maps. Bai et al. (2023) argue that since the generator weights are updated only with gradients derived from the discriminator, the discriminator 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, the discriminator must extract features the generator can decode to reconstruct the input.

Another potential solution is integrating GANs with methods based on explainable artificial intelligence (XAI) in the loss function. In a classification task, XAI provides explanations that show which parts of the input were most crucial in assigning the input to a label. These explanations can be derived using the model's gradients (Nielsen et al., 2022). One commonly used method is Saliency (Simonyan; Vedaldi; Zisserman, 2014), which calculates the gradients of the output concerning the input features to create explanations. These gradients help highlight areas where a slight change in the input would significantly alter the prediction. However, Saliency can sometimes produce noisy explanations. The Gradient$\odot$Input method (Shrikumar; Greenside; Kundaje, 2017) addresses this by multiplying the gradients with the input elementwise. It acts as a model-independent filter, reducing noise and smoothing the explanations. Another widespread method is DeepLIFT (Shrikumar; Greenside; Kundaje, 2017), which assigns importance to input features by comparing the activations caused by the actual input and a reference input in each neuron. The difference between the activations is used as importance scores for the input features. Gradient-based XAI produces detailed, fine-grained explanations at the pixel level with significant computational efficiency compared to XAI based on feature perturbation (Ribeiro; Singh; Guestrin, 2016; Lundberg; Lee, 2017). Therefore, it is necessary to conduct an in-depth study to investigate the potential of including the gradient-based methods in the loss function of GAN architectures to improve the generation of high-resolution artificial images.

1.1 MOTIVATIONS

The classification of medical images, such as histopathological images, is a challenging task due to the limited availability of labeled data and class imbalance (Ataky et al., 2020; Garcea et al., 2023; Guerrero et al., 2024). The low availability occurs due to the difficulty of obtaining tissue samples, the need for specialists to label the images, and the lack of standardization in the dimensions and formats of the images. Class imbalance is a common problem in classification tasks, where one or more classes have a significantly larger number of samples than others. Imbalance can hinder the training of machine learning models, as the model may learn to classify the majority class with high accuracy but fail to classify the minority classes. Increasing the diversity of training data is an effective strategy to overcome these challenges.

Data augmentation is a technique that generates new training samples from existing ones,

increasing the diversity of the training data and improving the model's generalization. Data augmentation is beneficial in image classification tasks, where the diversity of the training data is crucial for the model's performance. Classical data augmentation methods, such as rotations, flips, crops, and brightness adjustments, have significant limitations. They do not create new information but only modify the existing content, resulting in limited data diversity. In many cases, these transformations can introduce noise, remove important features, or be irrelevant to the specific domain. Additionally, they cannot model complex semantic variations, such as anatomical or pathological changes, and often lack robust randomness, producing augmented data with little variability. This can limit their effectiveness in reducing overfitting and improving the model's generalization.

The use of GANs in data augmentation offers significant advantages over classical methods. GANs can generate highly realistic and diverse synthetic data, going beyond simple geometric or brightness transformations to create new semantic variations that reflect complex patterns found in the real world. This capability is especially valuable in domains such as medical imaging, where it is crucial to represent rare anatomical or pathological variations. Additionally, the synthetic data generated by GANs can improve model generalization, reduce overfitting, and allow for the creation of balanced samples in imbalanced datasets. Finally, by directly modeling the data distribution, GANs can produce more representative and useful examples for classification.

Since Goodfellow et al. (2014) proposed the first GAN model, researchers have widely studied it in computer vision due to its data generation capabilities and the possibility of unsupervised representation learning. Diverse GAN architectures have provided significant contributions to different applications (Lorencin et al., 2021; Liao et al., 2022; Apostolopoulos; Papathanasiou; Panayiotakis, 2021; Ahmad et al., 2022; Sun et al., 2022; Çelik; Talu, 2022). However, applying GANs still presents significant challenges due to the imbalance between the G and D networks during training. The discriminator has an advantage over the generator, as it provides feedback to the generator based on the classification error. This feedback is essential for the generator to learn how to generate realistic images. However, the generator has no information on how the features of the images contribute to the classification. This lack of information makes the adversarial training unbalanced, leading to the generator failing to fool the discriminator even after the model converges (Wang et al., 2022; Bai et al., 2023).

Proposed solutions to the problems of training instability have been addressed in various works in the literature. However, these problems have no definitive solutions, making them open research issues. A potential solution is the integration of GANs with XAI-based methods in the loss function. XAI emerged from the need for more transparency and interpretability in the decision-making process of deep learning algorithms. The decision-making process of deep learning methods is known to be a black-box process, where there is knowledge about the input and output but no knowledge of why the transformation from one to the other occurred. XAI methods generate visualizations that illustrate which parts of an input were considered the most

important for the model, providing conditions for humans to understand why a decision was made. On the other hand, artificial neural networks simulate the functioning of the human brain, which includes the visual cortex responsible for processing visual information. From these facts arises the question: Are XAI-based methods capable of providing relevant information to a GAN as they provide it to humans? We aim to answer the following research questions: *Q1*: Which XAI-based method provides the most helpful information for training GANs? Moreover, *Q2*: Can an architecture combining GAN and an XAI-based method improve generation quality in high-dimensional image datasets?

1.2 OBJECTIVES

The primary objective of this research is to propose a novel training strategy that integrates XAI-based methods into the backpropagation algorithm of GANs to improve the generation of artificial images. To achieve this goal, we aim to:

1. Develop a loss function that incorporates gradient information from XAI-based methods to provide more informative feedback to the generator;
2. Evaluate the proposed model's performance on image datasets commonly used in the literature, such as MNIST and CIFAR10, in terms of image quality and diversity;
3. Perform data augmentation and classification tasks with state-of-the-art transformer models to assess the generalization capabilities of the proposed model in important histopathological image datasets;
4. Compare the proposed model's performance with traditional GAN architectures, measuring how much the proposal improves the performance of traditional GANs.

1.3 CONTRIBUTIONS

This research makes the following significant contributions:

1. An approach that feeds G with substantial information concerning the images' features, increasing the quality of the generated images;
2. The enrichment of D feedback that promotes a more significant variability in the generation of artificial images;
3. A new loss function that allow generating images with more realistic features, promoting an increase in the generalization and classification performance of state-of-the-art Transformer-based methods;
4. The indication of the best combination between GAN and XAI-based methods to generate and classify the histological images.

2 THEORETICAL FOUNDATION

This chapter presents the theoretical foundation of the research, covering the main concepts and techniques used in developing the proposed method. We introduce the concepts of histopathological images, computer-aided diagnosis, data augmentation, generative adversarial networks, and AutoDiff. We also discuss the principles of explainable artificial intelligence and transformer-based models, highlighting their relevance to the research.

2.1 HISTOLOGY IMAGES

Histopathological images are microscopic images of biological tissues commonly used in medical research and diagnostics. They help identify diseases by analyzing abnormalities in tissue morphology (Gurcan et al., 2009). The tissue samples are typically taken from a biopsy, stained with specific dyes such as Hematoxylin and Eosin (He et al., 2012), and captured using a microscope to reveal cellular structures and patterns. These images contain diverse types of cells, including epithelial cells, stromal cells, immune cells, and blood vessels, as well as extracellular matrix components like collagen and elastin fibers. The cellular and tissue structures in histopathological images provide valuable information about the underlying disease processes, making them essential for clinical diagnosis and research. Figure 1 shows examples of histopathological images of colon, breast, and liver tissue samples stained with Hematoxylin and Eosin.

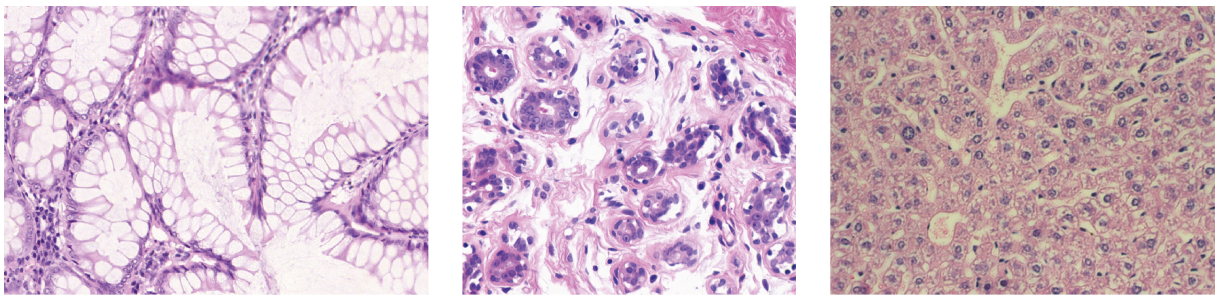


Figure 1 – Examples of a histopathological images of colon (a), breast (b), and liver tissue samples stained with Hematoxylin and Eosin.

Histopathological images are frequently used in digital pathology, which involves digitizing and analyzing tissue samples to help diagnose and treat diseases. Digital pathology allows pathologists to view and analyze tissue samples on a computer screen, enabling remote consultation, image sharing, and automated image analysis. The digitization of histopathological images has led to the development of computer-aided diagnosis systems that use machine learning algorithms to assist pathologists in interpreting and analyzing tissue samples. These systems can

help improve diagnostic accuracy, reduce workload, and provide valuable insights into disease progression and treatment response.

2.2 CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural networks are deep learning algorithms that simulate the functioning of the visual cortex of the animal brain to recognize and classify patterns in digital images (Yamashita et al., 2018; Sarvamangala; Kulkarni, 2022). The concept of CNN was first presented by Fukushima (1975) and gained significant repercussions after the victory of the proposal by Krizhevsky, Sutskever & Hinton (2017) in the ImageNet Large Scale Visual Recognition Challenge. The traditional architecture of a CNN consists of two main stages: pattern recognition and classification.

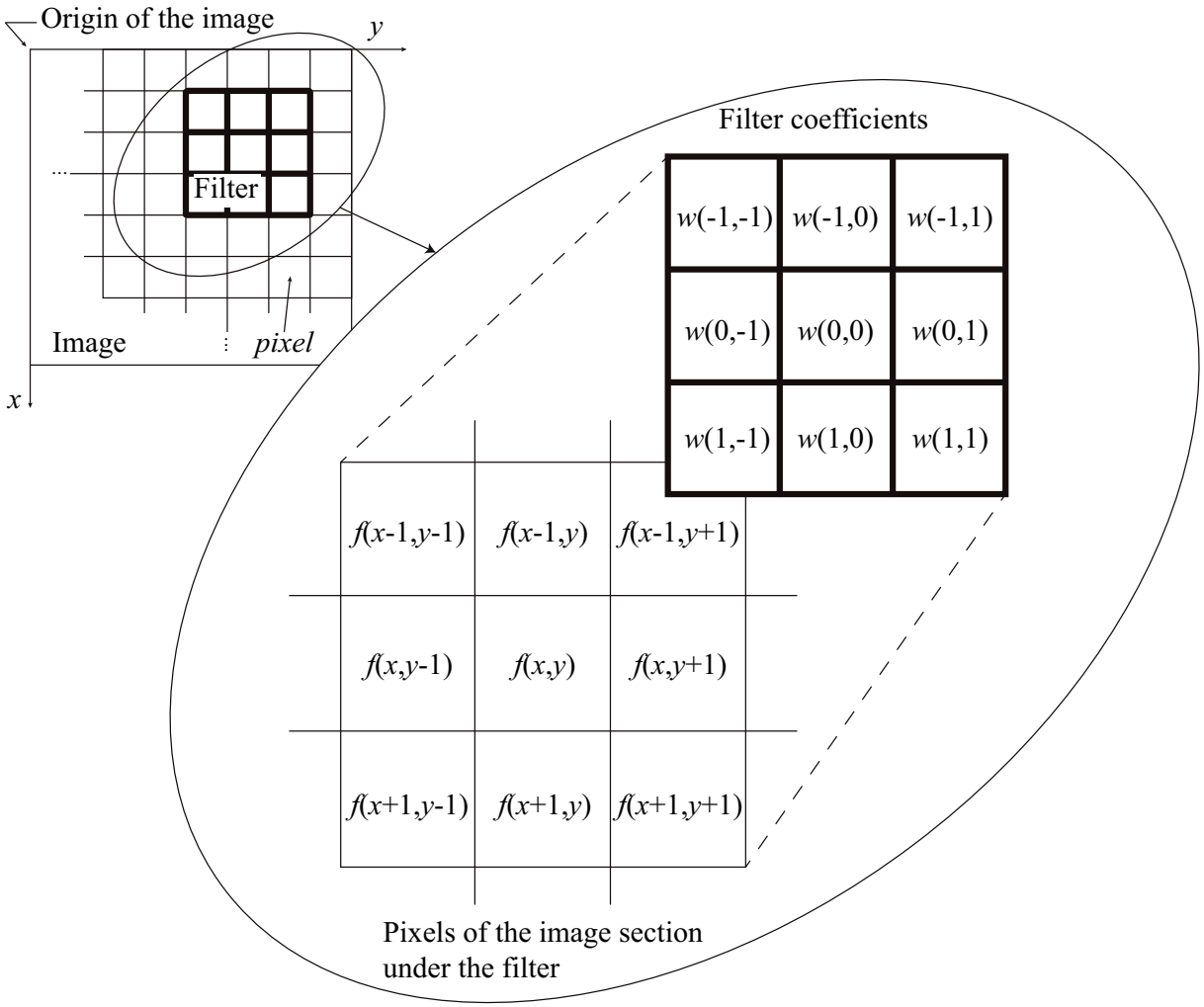
2.2.1 Pattern Recognition

The pattern recognition stage of a CNN consists of a series of convolutional layers that aim to extract features from input images. A convolutional layer is a set of various matrices, called spatial filters, that perform the convolution process. A spatial filter $w(x, y)$, also known as a kernel, is a small matrix of size $m \times m$ that stores coefficient values. Commonly, the value of m is an odd number to intuitively determine a central pixel (Gonzalez; Woods, 2017). The convolution, illustrated in Figure 2, involves moving the origin of the spatial filter, pixel by pixel, and performing operations between the filter coefficients and the pixels of I. The process starts at the top-left corner of I and advances sequentially in a horizontal scan, one line at a time. In the context of CNNs, the step size, in the number of pixels, that the filter advances in the image is called stride and is a natural number greater than zero. Mathematically, the convolution of a filter on an image is determined by summing the products of the filter coefficients, rotated by 180°, with the pixel values of the image within the filter's limits:

$$g(x, y) = w(x, y) * f(x, y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s, t) f(x - s, y - t), \tag{1}$$

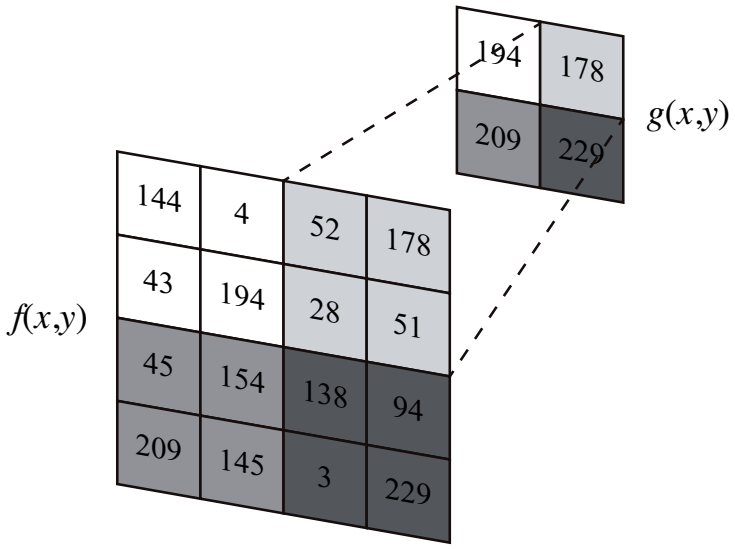
where a and b are positive integers, and x and y vary so that each pixel in w traverses all the pixels in f.

Each filter in a convolutional layer is responsible for extracting a different feature. In the context of CNNs, the resulting images from the convolutions are called feature maps. Commonly, feature maps undergo a dimensionality reduction process called pooling. Figure 3 exemplifies the most common type of pooling used in CNNs, max pooling. Similar to convolution, pooling is performed by sliding a filter over the image, following a stride value. The resulting image takes the maximum values from each filter. This procedure helps eliminate redundant information in the image, reduces the computational cost of feature extraction, and prevents overfitting (Liu et al., 2017; Li et al., 2022).



Source: Adapted from Gonzalez & Woods (2017).

Figure 2 – Schematic example of a spatial filter.



Source: Adapted from Yamashita et al. (2018).

Figure 3 – Example of max pooling operation with stride = 2.

2.2.2 Classification

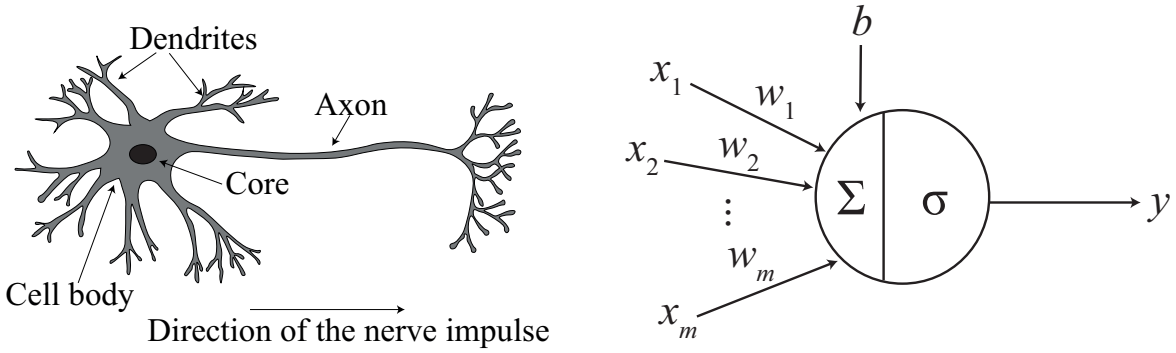
The second stage of the CNN aims to recognize patterns and classify images based on the features extracted by the convolutional layers. For this purpose, the feature maps resulting from the last layer of the feature extraction phase are transformed into a single one-dimensional vector, which is given as input to one or more fully connected (FC) layers.

The FC layers simulate a network of neurons in the animal brain. The biological neuron (Figure 4a) is a cell specialized in transmitting information and consists of three main parts: the cell body, dendrites, and axon. The direction of the nerve impulse, that is, the information, starts from the dendrites – where information from various neurons is received – goes to the cell body, and then to the axon, whose endings transmit the information to other neurons. This transmission is known as a synapse (Hall; Hall, 2020). To simulate the synapse mathematically, an artificial neuron (Figure 4b) receives an input vector $X = \{x_1, x_2, \cdots, x_m\}$, and each value x_i is weighted by an adjustable weight w_i. These values are summed with a bias value b and applied to an activation function σ to produce an output y, which is propagated to other artificial neurons. The artificial synapse is given by the equation:

$$y = \sigma(w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + b_i). \tag{2}$$

Commonly, the activation function used in CNNs is the Rectified Linear Unit (ReLU) because it allows reducing the training time by eliminating negative values. This function is calculated by:

$$\sigma(x) = \max(0, x). \tag{3}$$



(a) (b)

Source: (a) Adapted from Hall & Hall (2020), (b) Created by the author.

Figure 4 – Simplified illustration of the biological neuron (a) and the artificial neuron (b).

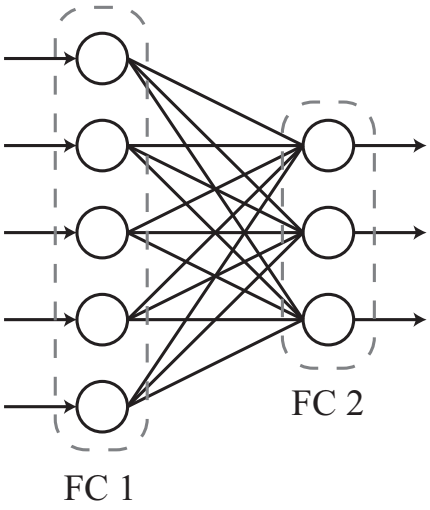
An FC layer composed of multiple artificial neurons is a function $f : \mathbb{R}^m \to \mathbb{R}^n$, that is, it maps an input X of dimension m to an output Y of dimension n. In a sequence of FC layers, the output of each artificial neuron is connected to the input of all neurons in the next layer, as illustrated in Figure 5. The output Y of an FC layer is given by:

$$Y = \begin{Bmatrix} y_1 = \sigma(w_{1,1}x_1 + w_{1,2}x_2 \cdots + w_{1,m}x_m) \\ \vdots \\ y_n = \sigma(w_{n,1}x_1 + w_{n,2}x_2 \cdots + w_{n,m}x_m) \end{Bmatrix} \tag{4}$$

The weights of the connections between neurons indicate the relevance of the connection to the classification process. These weights are adjusted during the network's training phase. Generally, the number of artificial neurons in the last FC layer is equal to the number of classes in the image dataset under study. In multi-class classification problems, the outputs are numbers in the range $[0, 1]$ that indicate the probabilities of an input image belonging to each class (Yamashita et al., 2018). The probabilities are obtained through the *softmax* activation function, which is given by:

$$\sigma(u_j) = \frac{e^{u_j}}{\sum_{k=1}^{M} e^{u_k}}, \tag{5}$$

where u_j is the output artificial neuron representing class j; and N is the total number of classes in the dataset under study. The class prediction of the CNN is given by the class with the highest probability.



Source: Created by the author.

Figure 5 – Illustrative diagram of connections between two FC layers.

2.2.3 Training a CNN

Training a CNN is a two-phase process that identifies the kernels in the convolutional layers, the weights, and the biases in the FC layers, which minimize the difference between the prediction made by the CNN and the true class of an image in the training dataset.

The first phase of training is called forward propagation and consists of obtaining the probabilities of the input image belonging to each of the classes in the dataset under study. From the obtained probabilities, the prediction error made by the CNN concerning the true class of

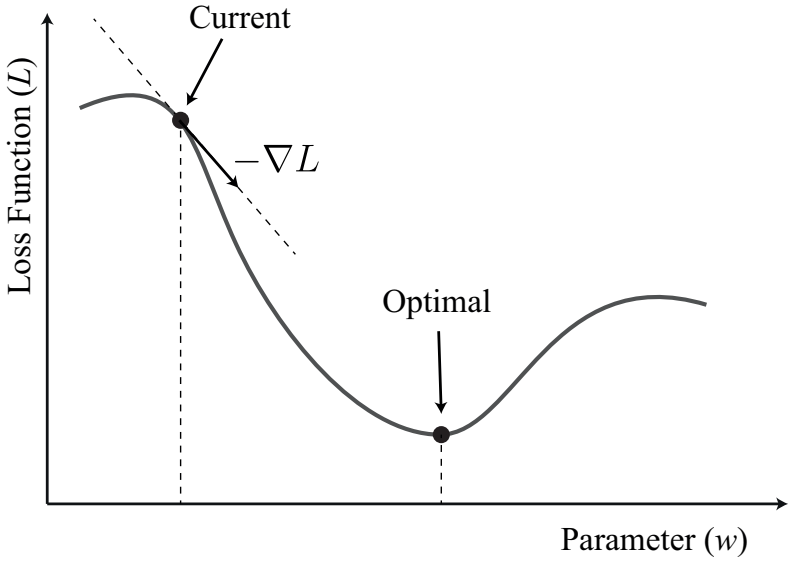
the image is calculated. The error is calculated using a function called the loss function (L). A commonly used function is the mean squared error (MSE), calculated by the equation:

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{6}$$

where y_i are the probabilities predicted by the CNN and $\hat{y}$ is a vector indicating the class of the image, with a value of 1 in the true class and zeros in the others.

The loss function indicates how accurate the prediction provided by the filters, weights, and biases of the CNN is. The lower the error value, the better the prediction made by the CNN. Therefore, training a CNN is an optimization problem where the goal is to minimize the loss function. This optimization is performed using the gradient descent algorithm, which provides the direction and magnitude in which the function has the steepest descent and which parameters should be updated so that the current error moves towards the optimal error. The gradient descent is calculated through the partial derivative of the loss function for each adjustable parameter present in the CNN. Equation 7 illustrates how the gradient descent is calculated for a hypothetical parameter w. The value of w is updated by its current value subtracted by a value α, which indicates the step size, times the result of the partial derivative of L with respect to w. Figure 6 graphically illustrates the process described by Equation 7.

$$w := w - \alpha\frac{\partial\mathcal{L}}{\partial w} \tag{7}$$



Source: Adapted from Yamashita et al. (2018).

Figure 6 – Illustrative diagram of connections between two FC layers.

The second phase of training is called backpropagation and consists of updating each parameter value of the CNN using the values obtained via gradient descent. Since a CNN is composed of multiple aligned layers, and the value of a parameter is a function of parameter

values from previous layers (Equation 2), the parameter updates are performed using the chain rule, which allows solving derivatives of composite functions.

The backpropagation algorithm is a recursive process that starts from the last layer of the CNN and advances sequentially to the first layer. The algorithm calculates the partial derivative of the loss function concerning each parameter in the CNN. The partial derivative of the loss function concerning the weights of the last FC layer is calculated by:

$$\frac{\partial \mathcal{L}}{\partial w_{i,j}} = \frac{\partial \mathcal{L}}{\partial y_i} \frac{\partial y_i}{\partial u_i} \frac{\partial u_i}{\partial w_{i,j}}, \tag{8}$$

where u_i is the output of the artificial neuron i in the last FC layer. The partial derivative of the loss function concerning the weights of the other layers is calculated similarly. The partial derivatives of the loss function concerning the biases are calculated in the same way. The backpropagation algorithm updates the weights and biases of the CNN using the values obtained by the partial derivatives. The algorithm iterates through the training dataset until the loss function converges to a minimum value, indicating that the CNN has learned the features and patterns present in the images.

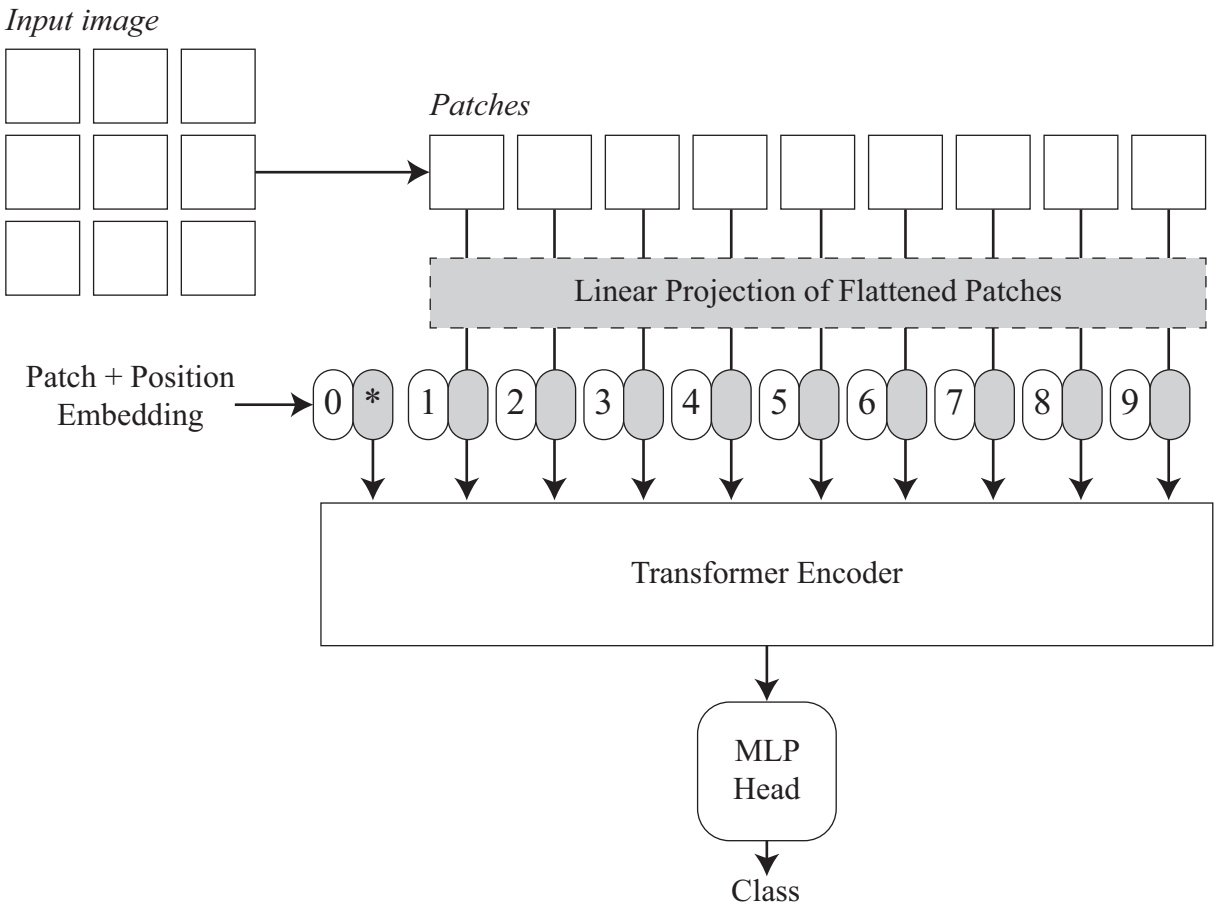
2.3 VISION TRANSFORMERS

Vision Transformer (ViT) (Dosovitskiy et al., 2021) is a deep-learning architecture for image recognition tasks. Figure 7 shows a schematic illustration of the ViT architecture. ViTs adapt the transformer architecture to process image data by dividing the input image into fixed-size patches and flattening them into sequences of vectors. These patches are then embedded into lower-dimensional representations and processed by the transformer encoder. The self-attention mechanism in the transformer allows the model to capture spatial relationships between image patches and learn hierarchical representations of image features.

ViTs have several advantages over traditional CNNs, such as capturing global context information, handling variable inputs, and scaling to larger image resolutions. They have been shown to achieve competitive performance on image classification benchmarks, such as ImageNet, and have been successfully applied to various computer vision tasks (Parvaiz et al., 2023). The transformer architecture's flexibility and scalability make ViTs a promising approach for developing CAD systems that can assist pathologists in diagnosing diseases and analyzing histopathological images (He et al., 2023).

2.4 COMPUTER-AIDED DIAGNOSIS

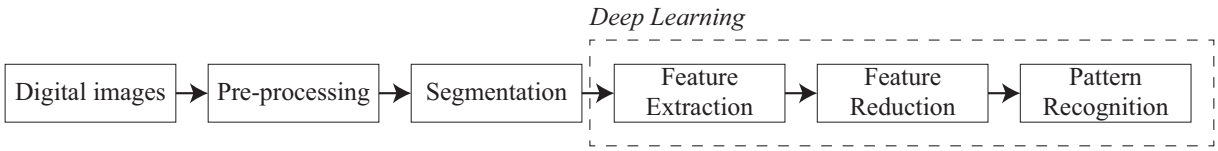
Computer-aided diagnosis (CAD) systems are software tools that help healthcare professionals interpret medical images and make diagnostic decisions. They are typically based on machine learning models that can learn complex patterns and relationships from datasets of annotated



Source: Adapted from Dosovitskiy et al. (2021).

Figure 7 – Schematic illustration of the ViT architecture.

images, providing additional information, highlighting crucial features, and reducing the risk of human error. Figure 8 shows a schematic illustration of the steps involved in a CAD system.



Source: Adapted from Jothi & Rajam (2017).

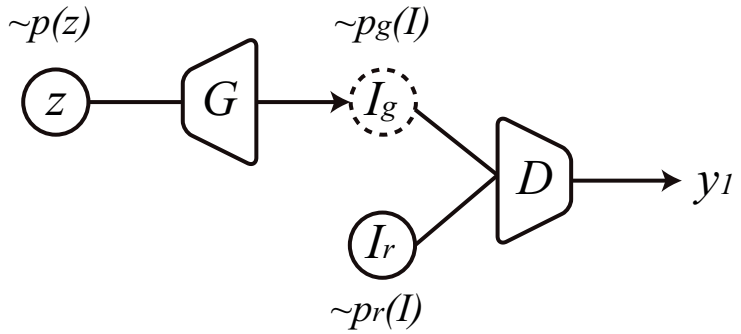
Figure 8 – Schematic illustration of a CAD system.

Deep learning algorithms, such as CNNs and transformers, have revolutionized the field of CAD by enabling the automatic extraction of features and the development of highly accurate diagnostic models. They can perform the three last steps of the CAD system —feature extraction, feature reduction, and pattern recognition— automatically, without the need for manual feature engineering. Transformers models are particularly well-suited for analyzing complex and high-dimensional data, such as histopathological images because they can learn hierarchical representations of image features. By training transformers on large datasets of annotated medical images, CAD systems can learn to recognize patterns and abnormalities associated with specific

diseases, making accurate predictions and assisting healthcare professionals in diagnosing and treating patients (He et al., 2023).

2.5 GENERATIVE ADVERSARIAL NETWORKS

Generative adversarial networks (GANs) are models that enable the generation of samples from a given distribution without the need to model the underlying probability density function explicitly (Yi; Walia; Babyn, 2019). GANs consist of two neural networks trained simultaneously: a generator (G) creates samples, and a discriminative network (D) categorizes the samples as authentic or artificial. The first GAN, the *vanilla*-GAN, was proposed by Goodfellow et al. (2014) to generate low-resolution images of handwritten digits, faces, and objects from a random signal. A schematic illustration exemplifying how the *vanilla*-GAN works is shown in Figure 9. G receives a random signal z from a Gaussian or uniform distribution p as input and produces the output image I_g, which belongs to a distribution p_g. The aim is for I_g to resemble I_r visually, meaning that the distribution p_g should be close to the real distribution p_r. D takes in a real or synthesized sample as input and outputs a value y_1 representing the probability that the input is real or generated by G. In vanilla-GAN, G and D are multilayer perceptrons, which are neural networks composed only of fully connected layers.



Source: Adapted from Yi, Walia & Babyn (2019).

Figure 9 – Schematic illustration of the *vanilla*-GAN.

The training of a GAN is based on game theory. The networks G and D act as adversaries competing with each other. D aims to maximize the probability of correctly classifying a sample, while G aims to minimize this probability. The desired outcome of this competition is the Nash equilibrium (Nash, 1950), where neither player has anything to gain by changing their strategy alone. Goodfellow et al. (2014) uses the analogy of a counterfeiter *versus* a police officer to illustrate the GAN training. In this analogy, G acts as a counterfeiter who produces fake products, and D acts as a police officer trying to determine whether the products are original. As the training progresses, the detective and the counterfeiter become increasingly proficient at their roles. The Nash equilibrium occurs when the police officer can not distinguish between original and counterfeit products, meaning the network D classifies with a 50% accuracy rate.

The GAN training process consists of two alternating stages. In the first stage, real and artificial images are fed into D. D updates its weights through backpropagation, while G remains unchanged. In the second stage, D is held constant, and G is updated using the gradients obtained from D. Mathematically, the training is a two-player minimax game in which G tries to minimize a value function $V(D, G)$, and D tries to maximize it:

$$\min_G \max_D V(D, G) = E_I[\log D(I_r)] + E_z[\log(1 - D(G(z)))], \tag{9}$$

where E_I and E_z are the expected values for all real and artificial instances, respectively; $D(I_r)$ is the probability of x_r belonging to the distribution p_g; $G(z)$ is the output of the network G given a random signal z; and $D(G(z))$ is the probability of an artificial instance being classified as real.

2.6 AUTODIFF

Automatic differentiation, or AutoDiff, is a set of techniques that efficiently enables computing partial derivatives of mathematical expressions. It is a crucial component in the backpropagation algorithm since it allows the automatic computation of the loss function's gradients concerning the model parameters. Autodiff is composed of two primary steps: forward pass and backward pass.

The forward pass creates a graph from input to output neurons. For each neuron, it computes the function value and its derivative concerning the input variables. Thus, considering a neuron i as a function f_i, and $x_1, x_2, \cdots, x_n$ as input variables, the AutoDiff computes and stores the value of the function $v_i = f_i(x_1, x_2, \cdots, x_n)$ in the graph and the partial derivative of v_i with respect to its inputs $\frac{\partial v_i}{\partial x_j}$.

The backward pass, or backpropagation, involves computing the gradients of the output variables concerning the input variables. The AutoDiff constructs the Jacobian matrix J by stacking the partial derivatives for each output concerning each input variable. Therefore, 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 Jacobian matrix J is defined as:

$$J = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}. \tag{10}$$

Finally, the chain rule is applied iteratively to compute the gradients for each neuron by multiplying the Jacobian matrix by the partial derivatives stored in the graph (Goodfellow; Bengio; Courville, 2016):

$$\frac{\partial y}{\partial x_j} = \sum_i \left(\frac{\partial y}{\partial v_i} \cdot \frac{\partial v_i}{\partial x_j} \right). \tag{11}$$

2.7 DATA AUGMENTATION

Data augmentation is a technique used to expand the size and diversity of training datasets by adding synthetic samples to the original data samples. It is commonly used in machine learning and deep learning to improve model generalization, prevent overfitting, and enhance model performance. The overfitting problem occurs when a model learns the training data too well, capturing the underlying patterns, noise, and irrelevant details. This results in the model performing exceptionally well on the training data but poorly on new, unseen data, as it struggles to generalize. Overfitting typically happens when a model is too complex relative to the training data available, such as when using a deep learning model on a small dataset. By creating new data samples from existing ones, data augmentation exposes the model to a broader range of variations and patterns in the data, making it more robust and adaptable to different scenarios (Mumuni; Mumuni, 2022).

Performing data augmentation via GANs is particularly important in medical imaging because it allows the creation of highly realistic synthetic images that can supplement limited or imbalanced datasets. Medical data is often scarce due to privacy concerns, high annotation costs, and the difficulty of collecting rare disease cases. GANs can generate synthetic medical images that resemble patient data, capturing complex patterns that traditional augmentation methods cannot produce (Mumuni; Mumuni, 2022; Kebaili; Lapuyade-Lahorgue; Ruan, 2023).

2.8 EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable artificial intelligence (XAI) is a field of research that aims to develop machine learning models capable of providing human-interpretable explanations for their predictions and decisions. XAI methods help users understand how a model works, why it makes specific predictions, and what features are most influential in the decision-making process. This transparency is essential for building trust in AI systems, ensuring accountability, and enabling users to validate and interpret the model's outputs (Minh et al., 2022).

The functioning of XAI methods is based on interpretability techniques that generate visualizations and explanations to facilitate understanding of the model. These techniques can be divided into three main categories: intrinsic interpretability, post-hoc interpretability, and interactive interpretability (Carvalho; Pereira; Cardoso, 2019; Minh et al., 2022). Intrinsic interpretability refers to models that are inherently interpretable, such as decision trees and linear regression (Carvalho; Pereira; Cardoso, 2019). These models are easy to interpret because their decisions are based on simple and transparent rules. Post-hoc interpretability involves applying interpretability methods to complex models, such as neural networks and random forests, to

generate understandable explanations. These methods include attribution-based interpretability techniques, such as LIME (Ribeiro; Singh; Guestrin, 2016), Saliency (Simonyan; Vedaldi; Zisserman, 2014), and DeepLIFT (Shrikumar; Greenside; Kundaje, 2017), which assign importance to specific features of the model. Interactive interpretability involves user interaction with the model to explore and understand its decisions. These methods include interactive visualizations, such activation maps, which allow users to investigate and analyze the model's behavior in real-time (Minh et al., 2022).

2.9 EVALUATION OF ARTIFICIAL IMAGE QUALITY

The goal of a GAN is to generate data in a distribution that matches the distribution of the original data. There are two ways to evaluate the correspondence of distributions. The first is the pixel distance calculation approach, which is considered naive and unreliable (Borji, 2019). The second approach is to calculate distances between features extracted from the original and artificial images, which provides a more stable and reliable evaluation since a feature vector is a high-level representation of an image. The most common feature extraction methods are the Inception Score (IS) (Salimans et al., 2016) and the Frechet Inception Distance (FID) (Heusel et al., 2017).

The Inception Score (IS) evaluates the quality and diversity of the generated images. It is calculated by feeding the generated images to a pre-trained Inception network and computing the KL divergence between the class distribution of the generated images and the overall class distribution. The IS is given by:

$$IS = \exp(E_x[D_{KL}(p(y|x)||p(y))]), \tag{12}$$

where $p(y|x)$ is the class distribution of the generated images, $p(y)$ is the overall class distribution, and D_{KL} is the Kullback-Leibler divergence. A higher IS indicates better image quality and diversity.

The Frechet Inception Distance (FID) measures the similarity between the feature distributions of the generated and real images. It is calculated by extracting features from the Inception network and computing the Frechet distance between the feature distributions. The FID is given by:

$$FID = ||\mu_x - \mu_y||^2 + \text{Tr}(\Sigma_x + \Sigma_y - 2(\Sigma_x\Sigma_y)^{1/2}), \tag{13}$$

where μ_x and μ_y are the mean feature vectors of the generated and real images, respectively, and Σ_x and Σ_y are the covariance matrices of the feature distributions. A lower FID indicates better image quality and similarity to the real images.

Both the IS and FID are widely used to evaluate the performance of GANs in generating high-quality images that resemble the original data distribution. These metrics provide quantitative

measures of image quality, diversity, and similarity to the real data, enabling researchers to compare different GAN models and assess their performance in generating realistic images (Souza et al., 2023).

2.10 CLASSIFICATION PERFORMANCE EVALUATION

In a supervised classification process, the correct classes, or labels, of the images are known beforehand. This allows measuring how accurate a classifier is in predicting labels. In the context of pattern recognition in images, the commonly used performance evaluation metric is accuracy.

Accuracy is a widely used measure because it allows for an intuitive interpretation of the results, when the datasets are balanced. Accuracy (Acc) is defined as the ratio of correctly classified cases to the total number of cases investigated, as indicated in Equation 14.

$$Acc = \frac{VP + VN}{CP + CN}, \tag{14}$$

where VP is the number of true positive cases, VN is the number of true negative cases, CP is the number of positive cases, and CN is the number of negative cases.

3 RELATED WORK

This chapter presents the most relevant works in the literature related to GANs, XAI, and Transformer-based models. We discuss the main contributions of each work and how they have influenced the development of these fields. We also highlight the limitations of existing methods and the literature gaps that this work aims to address.

3.1 GENERATIVE ADVERSARIAL NETWORKS

Goodfellow et al. (2014) introduced the first GAN architecture to generate images from a random signal based on features learned by G. However, the D and G architectures, multilayer perceptron-based networks, could not handle high-resolution image generation. Introducing convolutional-based GAN architectures made high-resolution image generation possible but also brought about training challenges. Studies in the literature traditionally aimed to improve GAN training by modifying the discriminator. For instance, the DCGAN model was the first to utilize convolutional layers in both D and G (Wang; She; Ward, 2021). The authors focused on improving the stability of the discriminator to enhance the training process. They incorporated batch normalization and leaky ReLU activations between intermediate layers of D and minimized the number of fully connected layers. The DCGAN employs the original Jensen-Shannon divergence as the loss function, but this approach can lead to mode collapse (Wang; She; Ward, 2021; Souza et al., 2023).

WGAN (Arjovsky; Chintala; Bottou, 2017) is a technique that addresses these problems by replacing the Jensen-Shannon divergence with the Wasserstein distance. This distance provides a continuous measure of the difference between probability distributions, which helps improve training stability. WGAN enforces the Lipschitz constraint on the discriminator by using weight clipping. However, this technique can lead to undesired side effects, such as vanishing or exploding gradients, which may decrease the model's learning capacity. WGAN-GP (Gulrajani et al., 2017), a significant improvement over WGAN, employs a gradient penalty term in the loss function instead of weight clipping to address these issues. This approach ensures that the model's learning capacity is not compromised, making it a more stable alternative to weight clipping.

RAGAN (Jolicoeur-Martineau, 2019) offers a new strategy to enhance training. It introduces the relativistic discriminator, which goes beyond simply classifying whether an input is real or fake. Instead, it estimates the probability that an authentic sample is more realistic than a fake sample and vice versa. This approach improves training stability and the quality of generated samples in specific scenarios. It provides a more detailed signal to the generator, helping it understand how to produce plausible samples independently and concerning the actual data distribution.

Other attempts to stabilize GAN training can also be cited. For example, the conditional GAN (cGAN) method (Wang; She; Ward, 2021), where a variable c is used as an additional input for G and D, allowing the generation process to be guided for different classes. Another example is the auxiliary classifier GAN (AC-GAN), proposed by Odena, Olah & Shlens (2017), where the softmax function was added to the output of the D network, allowing predictions about which class the input belongs to, in addition to predicting whether the sample is real or artificial. However, a problem indicated in these two approaches is that eventually, the class conditioning was lost during training, causing the GAN to ignore the class given as input. Chen et al. (2016) proposed a solution to this problem, the InfoGAN, where a third network, denoted by Q, was added. This network receives the artificial images and checks if the class information provided to G is present in the generated image.

Another strategy to condition the generation of images is to provide a real image from the dataset under study to the network G. This strategy is called domain translation. Domain translation involves transforming images belonging to one domain or class into another distinct domain. An example is transforming images of horses into images of zebras (Zhu et al., 2017). The main architecture used for this purpose is the cycle-consistent GAN (CycleGAN) proposed by Zhu et al. (2017). In CycleGAN, two generator models and two discriminator models are trained simultaneously. One of the generators receives images from the first domain as input and transforms them into images of the second domain. The other generator does the opposite: it receives images from the second domain and transforms them into images of the first domain. The discriminator networks determine how plausible the generated images are and update the generator networks. CycleGAN introduces a strategy that helps stabilize training, called the cycle-consistency loss function. This strategy involves using the output images of the first generator as input for the second generator, and vice versa. The output of one G model should match the image that served as input for the other, and the loss function then calculates the difference between the generated output of the second generator and the original image, and vice versa. This acts as a regularization of the generator models, guiding the image generation process in the new domain for image translation.

Lately, there has been a shift in focus towards enhancing the training of the image generator rather than the discriminator. The new methods involve providing additional information about the features of the images to G and striving for a more balanced competition between G and D. For example, Wang et al. (2022) introduced a training technique that improves the spatial awareness of G. This involves 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 to align the spatial awareness of G with D's attention maps. Bai et al. (2023) argue that as 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.

3.2 EXPLANABLE ARTIFICIAL INTELLIGENCE

Works that present XAI methods to provide visualizations in image classification systems can be divided into two main approaches: 1) methods that aim to explain a classification decision from an input, and; 2) methods that aim to interpret how the intermediate layers see the external world in general (Barredo Arrieta et al., 2020).

3.2.1 Methods that Explain a Classification Decision

In general, methods of the first approach work as follows: given an image dataset, an image from the test set is given as input to a pre-trained CNN. The XAI methods then attempt to explain the decision made by the CNN through a reverse mapping that starts from the output layer and passes through all the intermediate layers. This mapping aims to identify which regions of the input image caused the highest activations in the CNN. The result is a saliency map, which is a heat map where the hottest regions indicate the areas of the image that were most relevant to the decision-making process.

These methods can be categorized into methods based on feature perturbation (Ribeiro; Singh; Guestrin, 2016; Lundberg; Lee, 2017) and techniques based on gradient information (Simonyan; Vedaldi; Zisserman, 2014; Shrikumar; Greenside; Kundaje, 2017; Nielsen et al., 2022). The first type involves perturbing the input by masking or altering its values and observing the impact of these changes on the network's performance. For instance, the method LIME (Ribeiro; Singh; Guestrin, 2016) segments the image into superpixels and turns them on and off to generate a new set of images. It uses the network to predict the outcomes for the perturbed images, while a weighted linear model is trained to indicate the importance of each superpixel in the prediction. Another example is SHAP (Lundberg; Lee, 2017), which considers pixels or superpixels as features and examines all possible combinations of features, determining how prediction changes when features are added to different subsets of other features.

The main drawback of methods based on feature perturbation is that they require multiple passes through the network, making them computationally intensive. They also include potential instability, as small changes in data can lead to different interpretations. Gradient-based methods, on the other hand, directly use the internal workings of the network to estimate attribution scores. These techniques offer several benefits, such as computational efficiency and detailed, fine-grained explanations at the pixel level that allow a precise understanding of which parts of the input are most influential in the model's decisions (Nielsen et al., 2022).

A widely used method is the Saliency method (Simonyan; Vedaldi; Zisserman, 2014), which creates explanations by calculating the gradients of the network's output concerning the input features. It allows highlighting regions where a slight change in the input would significantly change the predictions. However, Saliency can produce noisy explanations. The Gradient$\odot$Input (Shrikumar; Greenside; Kundaje, 2017) 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 smoothens the explanations (Nielsen et al., 2022). Another widely used method is DeepLIFT (Shrikumar; Greenside; Kundaje, 2017), 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.

3.2.2 Methods that Interpret Intermediate Layers

Methods of the second approach, in turn, produce visualizations that aim to illustrate how convolutional layers interpret the external world. The construction of this type of visualization is carried out by extracting visual information stored during training and is not necessarily related to any specific input. Generally, these methods receive a signal as input and perform successive transformations to iteratively build the visualization. The result is an image that illustrates the visual information present in the neurons, channels, or internal layers of the CNN.

Mahendran & Vedaldi (2015), for example, proposed a method based on an inverse function focusing on illustrating the information present in the layers. The method is an iterative process that starts from random noise and transforms it into the representation using the activation information stored in the CNN. The authors concluded that convolutional layers could retain photographically accurate information about the images.

Nguyen et al. (2016) proposed an activation maximization method that aims to identify what each artificial neuron learned during training. The strategy used by the authors was to take a random signal as the initial solution and iteratively calculate, via backpropagation, how the color of each pixel in the image should be altered to increase the activation of a neuron. The proposed method revealed the characteristics learned by each neuron in an interpretable way, as it was able to generate images similar to real ones. This strategy was the same used in the deep dream method (Mordvintsev; Olah; Tyka, 2015) and other technologies developed by Google (Olah; Mordvintsev; Schubert, 2017; Olah et al., 2018) that allow visualizing characteristics stored in each neuron, channel, or layer.

3.3 TRANSFORMER-BASED MODELS

The advent of transformers in computer vision has led to substantial advancements compared to convolutional neural networks. These approaches can capture global contextual information and long-range dependencies in images, allowing more accurate understanding of complex visual patterns. As a result, they provide state-of-the-art performance in image classification. The Vision Transformer (ViT) model (Dosovitskiy et al., 2021) is a prime example. It divides images into fixed-size patches, then linearly embeds them and feeds them into a standard transformer architecture. It leverages self-attention mechanisms to capture relationships between these patches, allowing the model to understand the global structure of the image.

Wang et al. (2021b) introduced the Pyramid Vision Transformer (PVT) that extends the ViT model by incorporating a hierarchical structure operating with different patch 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. This approach enhances the model's ability to capture local and global features, improving its performance in image classification tasks.

The Data-efficient Image Transformer (DeiT) (Touvron et al., 2021) introduces a more data-efficient training strategy, enabling transformers to perform well 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.

The Convolutional Vision Transformer (CoAtNet) (Dai et al., 2021) is a hybrid architecture that combines the strengths of CNNs and Transformers. It integrates convolutional layers in the early stages of the network to extract local features, while later stages use transformers to model long-range dependencies and global context. 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. The hybrid approach allows CoAtNet to leverage the inductive biases of CNNs, such as translation invariance and locality, while also benefiting from the global attention mechanism of transformers.

3.4 LIMITATIONS AND CHALLENGES

Despite the significant advancements in GANs, several challenges remain to be addressed. GANs are known to suffer from training instability, mode collapse, and convergence issues, which can hinder the generation of high-quality images. Improving the stability and convergence of GANs is an ongoing research challenge, as it requires balancing the competition between the generator and discriminator networks and ensuring that the generator learns to generate diverse and realistic samples. The approaches proposed by Wang et al. (2022) and Bai et al. (2023) aim to address these challenges but a study addressing the inclusion of XAI-based methods in the loss function has yet to be performed.

Regarding the XAI methods available in the literature, we understand that methods belonging to the second approach, where visualizations illustrate the information contained in convolutional layers, are unsuitable for our proposal as they require a trained CNN to produce meaningful visualizations, making them incompatible with our intended purpose for these techniques in GAN training. Among the XAI methods of the first approach, feature perturbation-based methods (Ribeiro; Singh; Guestrin, 2016; Lundberg; Lee, 2017) are not suitable for our objectives, as they are computationally intensive due to the need for multiple passes through the network and potential instability. Therefore, gradient-based methods (Simonyan; Vedaldi; Zisserman, 2014;

Shrikumar; Greenside; Kundaje, 2017; Nielsen et al., 2022) are more computationally efficient and provide detailed and precise pixel-level explanations, making them the most suitable for our proposal.

4 METHODOLOGY

This chapter presents the proposed method, the XGAN, which combines GANs with XAI-based methods to generate high-quality artificial images. We describe the architecture of the XGAN model, the training process, and the evaluation metrics used to assess the performance of the model. We also discuss the datasets and experiments conducted to validate the proposed method and compare it with existing approaches.

4.1 METHOD OVERVIEW

A schematic summary of the proposed method's structure – the XGAN – is illustrated in Figure 10. It includes a generator (G) and a discriminator (D). The G network receives a random signal vector z and outputs an image $G(z)$. At the same time, D classifies original x and artificial images $G(z)$. The model uses XAI-based methods 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 gradient information (E) extracted by the XAI-based methods 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.

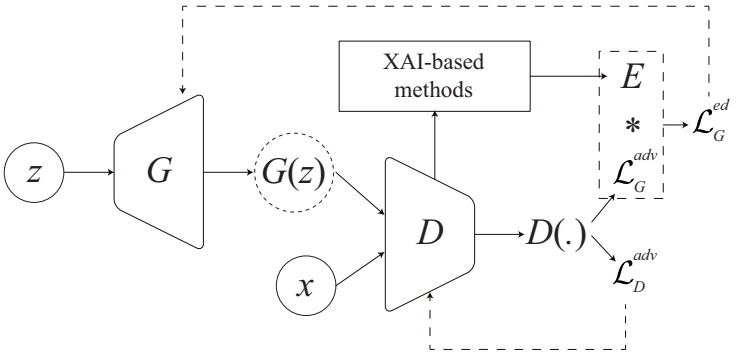


Figure 10 – Schematic summary of the proposed model.

The gradients indicate the adjustment needed in each weight of G to guide its loss function toward the optimum. Including E in the gradients emphasizes areas corresponding to objects of interest while reducing the influence of less relevant regions. Our approach involves a student-teacher dynamic. In this setup, E represents a test answer in which the professor (D) provides the student (G) with their test score, indicating which features closely resemble reality and which do not match the original images. Instead of propagating a single value that reflects D's

classification error, we propagate a matrix containing 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 J is the Jacobian matrix defined by Equation 10, and $\vec{v}$ is a multidimensional vector of the same dimension as the explanations E with 1 in all positions. 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 DCGAN, WGAN-GP, and RAGAN models to define the adversarial loss functions $\mathcal{L}_D^{adv}$ and $\mathcal{L}_G^{adv}$, and the gradient-based methods Saliency, DeepLIFT, and Gradient$\odot$Input to generate the explanations E.

4.2 ADVERSARIAL LOSS FUNCTIONS

In this section, we describe the adversarial loss functions used in the proposed XGAN model. We detail the loss functions for the DCGAN, WGAN-GP, and RAGAN architectures, which serve as the base models for our experiments.

4.2.1 DCGAN

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

$$\mathcal{L}_D^{\text{DCGAN}} = \mathbb{E}_{x\sim p(x)}[\log(D(x))] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))], \tag{3}$$

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

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

4.2.2 WGAN-GP

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

$$\mathcal{L}_D^{\text{WGAN-GP}} = -\mathcal{L}_W + \mathcal{L}_{GP}, \tag{5}$$

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

$$\mathcal{L}_W = \mathbb{E}_{x\sim p(x)}[D(x)] - \mathbb{E}_{z\sim p(z)}[D(G(z))], \tag{6}$$

and

$$\mathcal{L}_{GP} = \lambda \mathbb{E}_{\hat{x}\sim p(\hat{x})}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{7}$$

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:

$$\mathcal{L}_G^{\text{WGAN-GP}} = -\mathbb{E}_{z\sim p(z)}[D(G(z))]. \tag{8}$$

4.2.3 RAGAN

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

$$\mathcal{L}_D^{\text{RAGAN}} = \mathcal{L}_D^{\text{DCGAN}} + \mathcal{L}_{rel}, \tag{9}$$

where

$$\mathcal{L}_{rel} = -\frac{1}{2}\mathbb{E}_{x\sim p(x), z\sim p(z)}[\log(D(x) - D(G(z))]. \tag{10}$$

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

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

4.3 XAI-BASED METHODS

For this work, we used gradient-based techniques to extract the most critical features from D's gradients. We opted to use this type of method due to its computational efficiency and capacity for creating fine-grained pixel-level explanations (Nielsen et al., 2022). We used the Saliency, DeepLIFT, and Gradient$\odot$Input methods to generate the explanations E.

To calculate the Saliency (Simonyan; Vedaldi; Zisserman, 2014) explanation (E_{Saliency}) for a fake image $G(z)$, we calculated the partial derivative of associated output $D(G(z))$ concerning the input $G(z)$:

$$E_{\text{Saliency}} = \frac{\partial D(G(z))}{\partial G(z)}. \tag{12}$$

To determine the Gradient$\odot$Input (Shrikumar; Greenside; Kundaje, 2017) explanation ($E_{\text{Gradient}\odot\text{Input}}$), we calculated the elementwise multiplication of gradients by the input:

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

Finally, to define DeepLIFT (Shrikumar; Greenside; Kundaje, 2017), 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:

$$E_{\text{DeepLIFT}} = \sum_{i=1}^{n} C_{\Delta\eta_i \Delta t} = \Delta t, \tag{14}$$

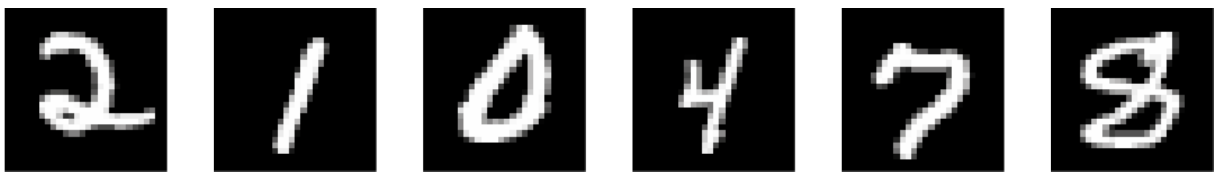
where $\Delta\eta_i$ is the difference between neuron activations caused by $G(z)$ and x^0, and $C_{\Delta\eta_i \Delta t}$ is the contribution score of $\Delta\eta_i$ to Δt.

4.4 DATASETS

We conducted experiments using the MNIST, CIFAR-10, Colorectal Cancer (CR), Liver Age (LA), Liver Gender (LG), and Breast cancer (UCSB) datasets.

4.4.1 MNIST and CIFAR-10

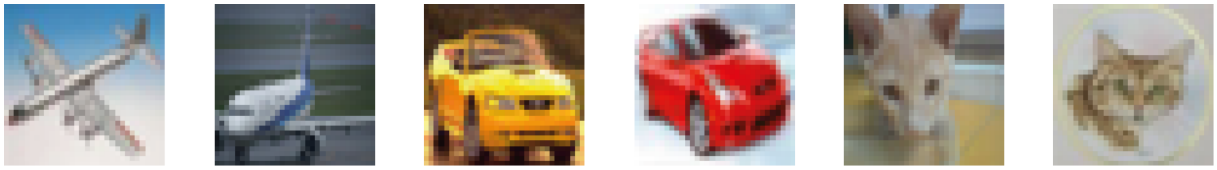
The MNIST dataset (Lecun et al., 1998) (Figure 11) is a set of 70,000 images of size 28×28 of handwritten digits. The digits were normalized for size and centered in fixed-size grayscale images.



Source: Lecun et al. (1998).

Figure 11 – Examples of samples from the MNIST dataset.

The CIFAR-10 dataset (Krizhevsky, 2009) (Figure 12) consists of 60,000 RGB images of size $32 \times 32 \times 3$. The dataset has 10 classes with everyday objects that include, for example, airplanes, automobiles, and cats. Each class contains 6,000 images.



Souce: Krizhevsky (2009).

Figure 12 – Examples of samples from the CIFAR-10 dataset of the classes airplane, automobile, and cat.

4.4.2 Histopathological Images

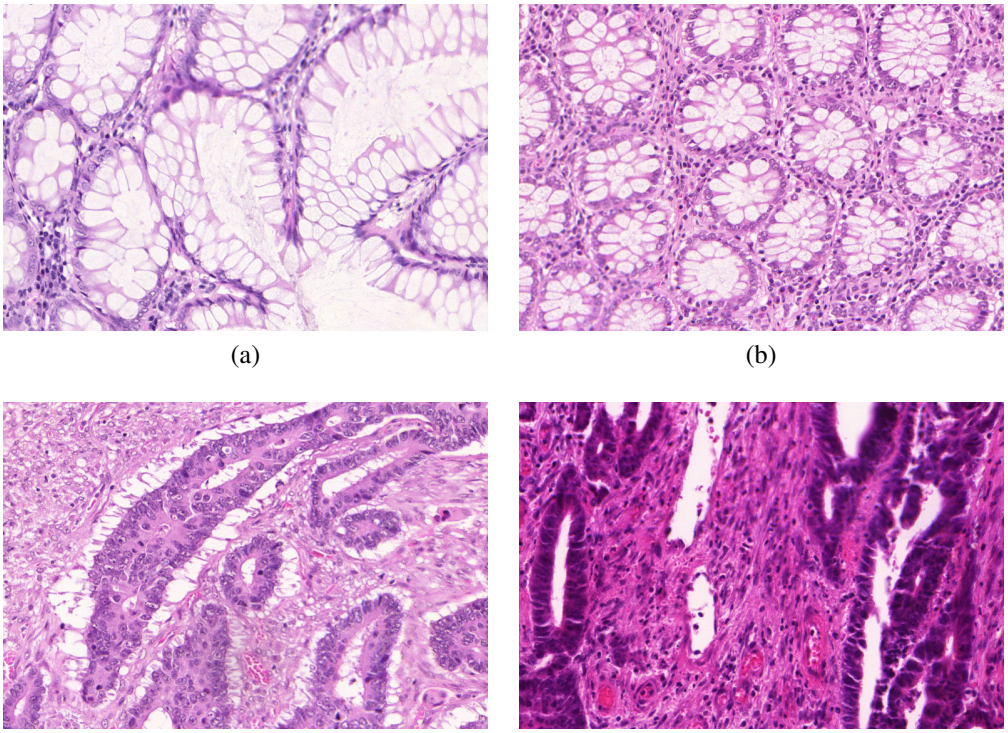
The field of histological imaging presents unique challenges due to two main issues: a deficit of labeled images and class imbalance. The limited availability of labeled images results from ethical concerns regarding patient privacy, industry standards in the medical field, and a lack of integration between medical information systems (Yi; Walia; Babyn, 2019). Furthermore, when labeled images are accessible, another challenge arises from the conditions of the image databases, including class imbalances and inconsistencies in dimensions and formats. These factors can compromise the effective training of classification systems and lead to overfitting (Madani et al., 2018; Frid-Adar et al., 2018).

4.4.2.1 Colorectal Cancer

The Colorectal Cancer (CR) dataset (Sirinukunwattana et al., 2017) (Figure 13) 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$\times$ magnification. The images have different sizes, ranging from 567×430 to 775×522 pixels.

4.4.2.2 Liver Tissue

The LA and LG datasets (Shamir et al., 2008) (Figure 14 and Figure 15) 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: one month (100), six 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$\times$ 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).



(a) (b) (c) (d)

Figure 13 – Example images from the CR dataset: (a) and (b) benign tumors, (c) and (d) malignant tumors.

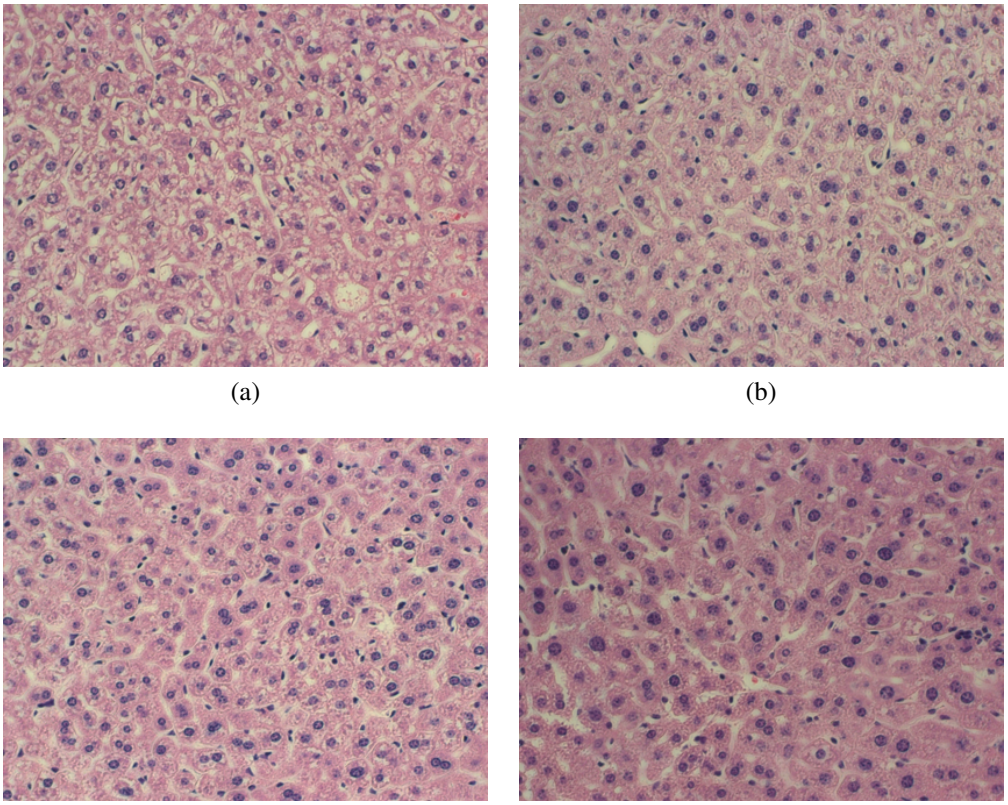
4.4.2.3 Breast Cancer

The UCSB dataset (Gelasca et al., 2008) (Figure 16) 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.

4.5 IMAGE QUALITY EVALUATION

We applied the FID metric (Heusel et al., 2017) 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:

$$FID = \|\mu_{\mathrm{r}} - \mu_{\mathrm{f}}\|^2 + \mathrm{Tr}(\Sigma_{\mathrm{r}} + \Sigma_{\mathrm{f}} - 2(\Sigma_{\mathrm{r}} \cdot \Sigma_{\mathrm{f}})^{0.5}), \tag{15}$$



(a) (b)

(c) (d)

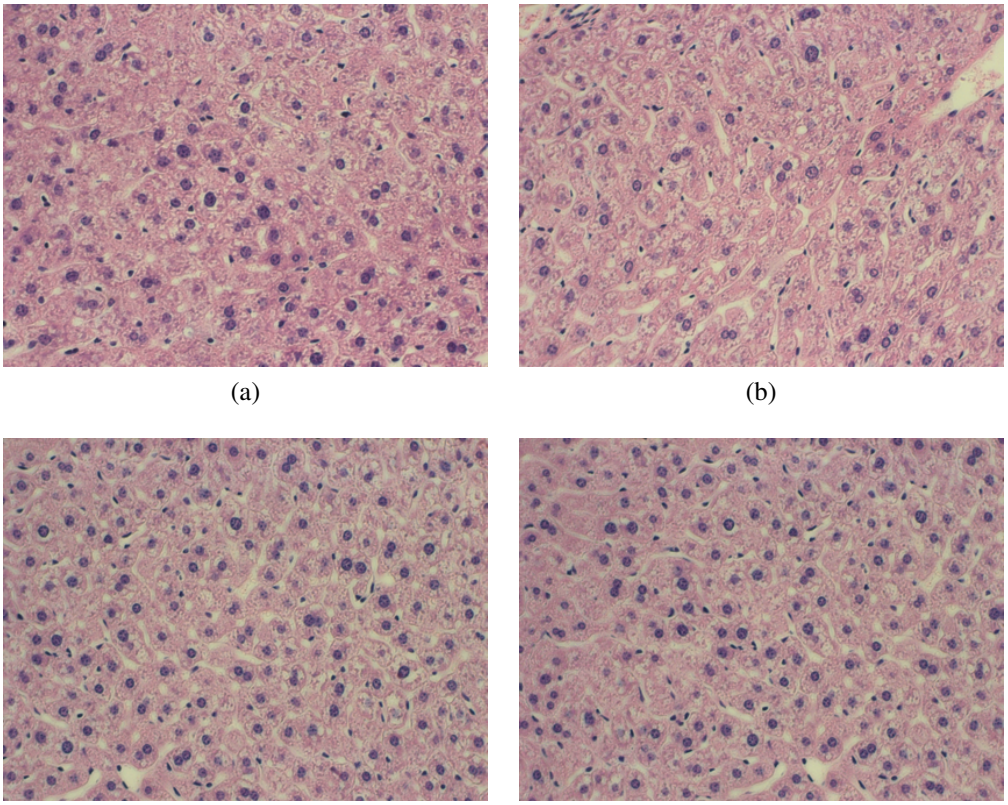
Figure 14 – Example images from the LA dataset: (a) one month, (b) six months, (c) 16 months, and (d) 24 months.

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 (Salimans et al., 2016) to estimate the diversity of 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:

$$IS = \exp\left(\mathbb{E}_x\left[D_{\mathrm{KL}}(p(y|x)||p(y))\right]\right), \tag{16}$$

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.



(a) (b)

(c) (d)

Figure 15 – Example images from the LG dataset: (a) male and (b) female.

4.6 CLASSIFICATION EVALUATION

To train and evaluate the GAN and XGAN models, we employed a strategy based on regions of interest (ROIs). Figure 17 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.

4.7 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,

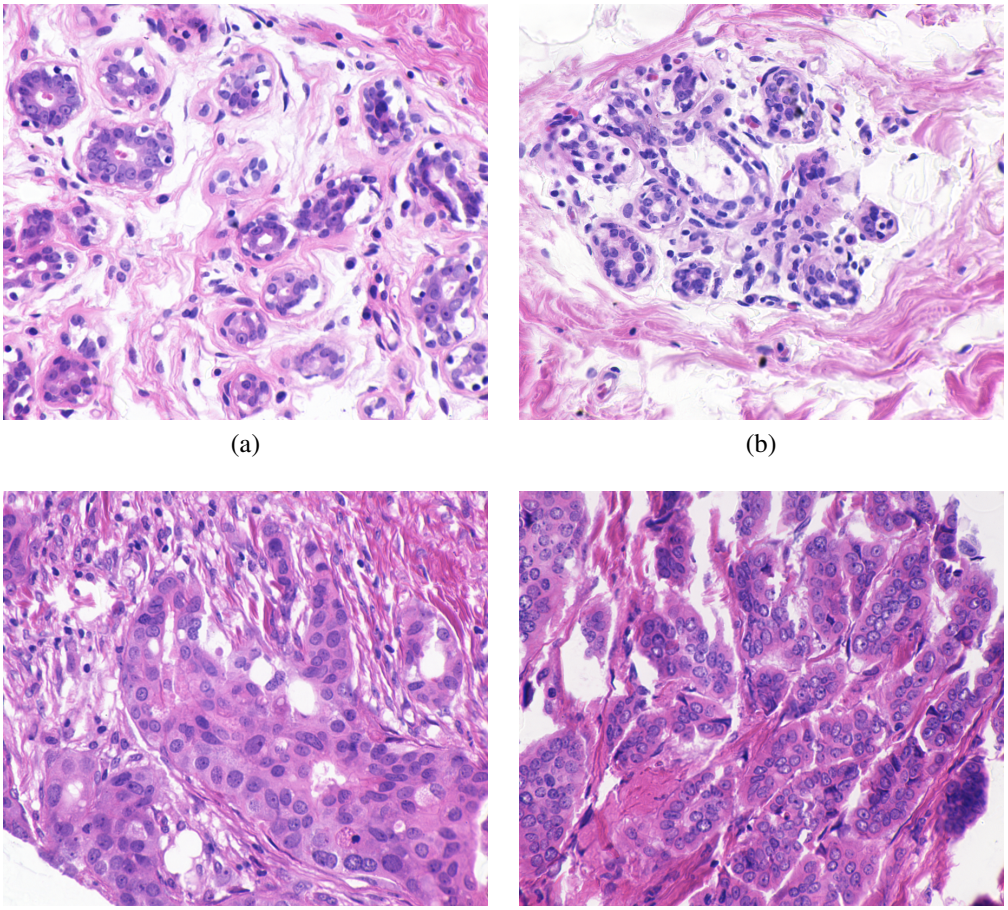


Figure 16 – Example images from the UCSB dataset: (a) and (b) benign tumors, (c) and (d) malignant tumors.

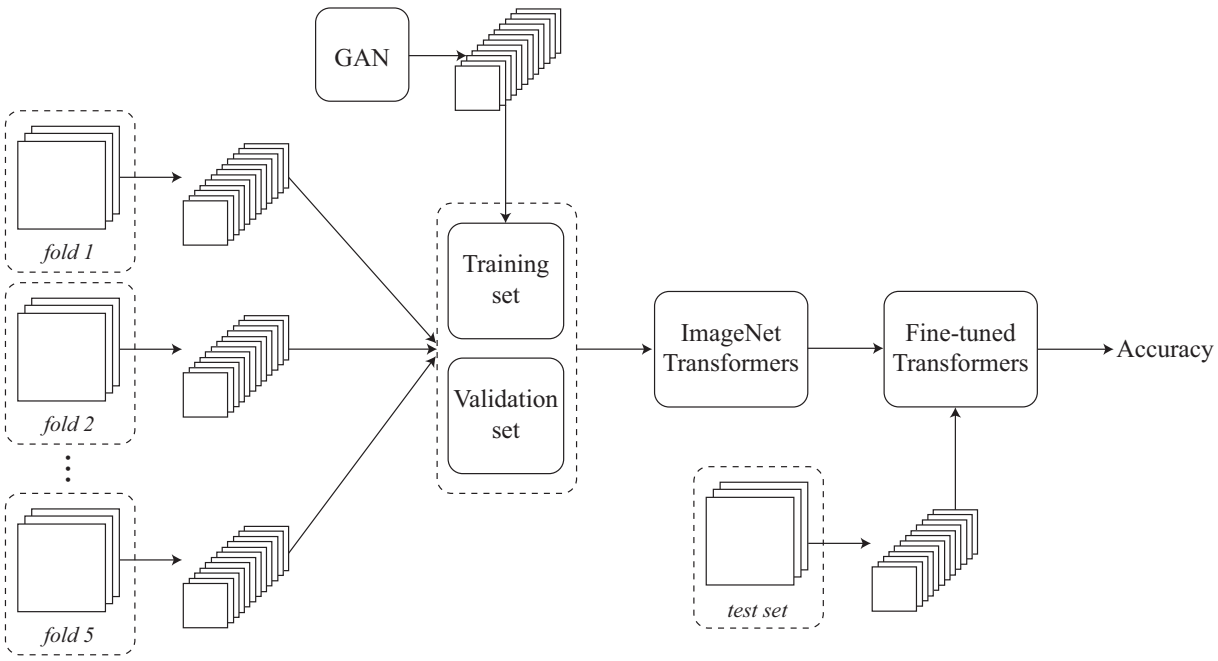


Figure 17 – Schematic illustration of the classification evaluation process.

2.10GHz, NVIDIA® GeForce RTX™3090 card, 64 GB of RAM and Windows operating system with 64-bit architecture.

5 RESULTS AND DISCUSSION

We conducted our experiments in two distinct directions. First, we evaluated the image generation capabilities of the proposed method using reference datasets such as MNIST and CIFAR10. The goal was to assess our method's image generation potential on simpler datasets commonly used in the literature. Training GANs on more complex datasets, such as histopathological images, can be computationally expensive. This preliminary testing allows for multiple training runs, enabling us to extract statistical measures and adjust parameters to preview the technique's performance. For this evaluation, we utilized FID and IS metrics to assess the generated images' quality quantitatively.

The second direction of our experiments was to evaluate the proposed method on histopathological datasets. We used four distinct datasets: CR, LA, LG, and UCSB. These datasets are more complex and have a higher resolution than MNIST and CIFAR10. We used the same metrics as in the first direction to quantitatively evaluate the generated images' quality. We also utilized the resulting artificial images to perform data augmentation and train transformer-based models for image classification tasks. The idea was to determine if the proposed method could provide more meaningful features than original GAN architectures. We used the accuracy metric to assess the classification performance of the models.

5.1 RESULTS ON MNIST AND CIFAR10

Table 1 shows the averaged FID and IS on MNIST and CIFAR10 datasets regarding GAN and XGAN architectures. 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. The green arrow indicates an improvement, while the red arrow indicates a decrease in the metric.

Regarding the results on the MNIST dataset, it is possible to note that using the Saliency method with the DCGAN architecture improved the quality of the artificial images. XDCGAN (Saliency) produced images of better quality, with an FID score around 3.72% lower than DCGAN. The images generated with XDCGAN (Saliency) were more diverse, evidenced by an IS score of 2.1294, slightly higher than the 2.1289 obtained with DCGAN. Considering the RAGAN architecture, the proposed method also provided relevant results. The highlight was XRAGAN (Saliency), which improved the quality and variability of artificial images. This combination made obtaining an FID of 15.79% lower than RAGAN and an IS of 2.0041 possible. The WGAN-GP architecture also benefited from the proposed method. The best results were obtained with XWGAN-GP (DeepLIFT) and XWGAN-GP (Gradient$\odot$Input), which provided FID scores up to 37.8% lower than WGAN-GP. The IS score was also higher, with XWGAN-GP

(Gradient⊙Input) achieving an IS of 2.2822, approximately 4.94% higher than WGAN-GP.

Considering the results on the CIFAR10 dataset, the proposed method also improved the quality of artificial images. Using the XAI methods with the DCGAN architecture, it was possible to obtain a slight quality improvement in the artificial images with the XDCGAN (Saliency) and XDCGAN (DeepLIFT) methods. These methods provided FIDs of 30.9560 and 30.8860, lower values than DCGAN (FID = 31.0575). The XDCGAN (Saliency) also increased the IS metric, 2.1724, against 2.1721 provided by DCGAN. On the other hand, when considering the RAGAN architecture, it was possible to obtain an increase in the quality of artificial images with all XRAGAN combinations, with emphasis on XRAGAN (DeepLIFT), which, in addition to promoting increased quality, also generated images with more variety compared to RAGAN. When considering the WGAN-GP architecture, we did not achieve an increase in quality when generating images from the CIFAR10 dataset. Despite this, it is possible to notice that in all XWGAN-GP combinations, there was a slight increase in the variability of the artificial images.

Table 1 – FID and IS scores for MNIST and CIFAR10 datasets.

		MNIST		CIFAR10	
		FID	IS	FID	IS
DCGAN	no XAI	2.3428	2.1289	31.0575	6.7349
	Saliency	**2.2572** ↓	**2.1294** ↑	30.9560 ↓	**6.7764** ↑
	DeepLIFT	2.6792 ↑	2.1239 ↓	**30.8860** ↓	6.6785 ↓
	Gradient⊙Input	2.7468 ↑	2.1168 ↓	31.1186 ↑	6.6900 ↓
RAGAN	no XAI	33.8299	1.9831	33.8299	1.9831
	Saliency	**28.8780** ↓	**2.0041** ↑	**28.8780** ↓	**2.0041** ↑
	DeepLIFT	33.5152 ↓	1.9532 ↓	33.5152 ↓	1.9532 ↓
	Gradient⊙Input	31.1702 ↓	1.9652 ↓	31.1702 ↓	1.9652 ↓
WGAN-GP	no XAI	11.0952	2.1721	33.9477	6.5254
	Saliency	12.0221 ↑	2.1724 ↑	35.3156 ↑	**6.6791** ↑
	DeepLIFT	**7.5686** ↓	2.2673 ↑	34.4661 ↑	6.6555 ↑
	Gradient⊙Input	7.6735 ↓	**2.2822** ↑	35.1785 ↑	6.5954 ↑

To illustrate the quantitative results, we present in Figure 18 some examples of artificial images generated by WGAN-GP and XWGAN-GP variants on the MNIST dataset. It is worth noting that images produced by WGAN-GP have some defects, such as noise and artifacts. These defects are highlighted with red arrows in Figure 18a. Meanwhile, the images generated by XWGAN-GP (Saliency) have a blur effect (Figure 18b). Although the blur effect is undesirable in generating artificial images, this indicates that information was included in the generator training. It is also possible to note that this combination eliminated the noise and artifacts. On the other hand, the images created by XWGAN-GP (DeepLIFT) and XWGAN-GP (Gradient⊙Input) (Figure 18c and Figure 18d) are crisp and noise-free, nearly indistinguishable from the images in the original dataset (as shown in Figure 11). Therefore, the difference in FID presented in Table 1 is due to these factors. These combinations resulted in the most significant reduction in FID, with

XWGAN-GP (DeepLIFT) showing a decrease of 37.8% and XWGAN-GP (Gradient⊙Input) showing a decrease of 30.9% in comparison to XGAN-GP.

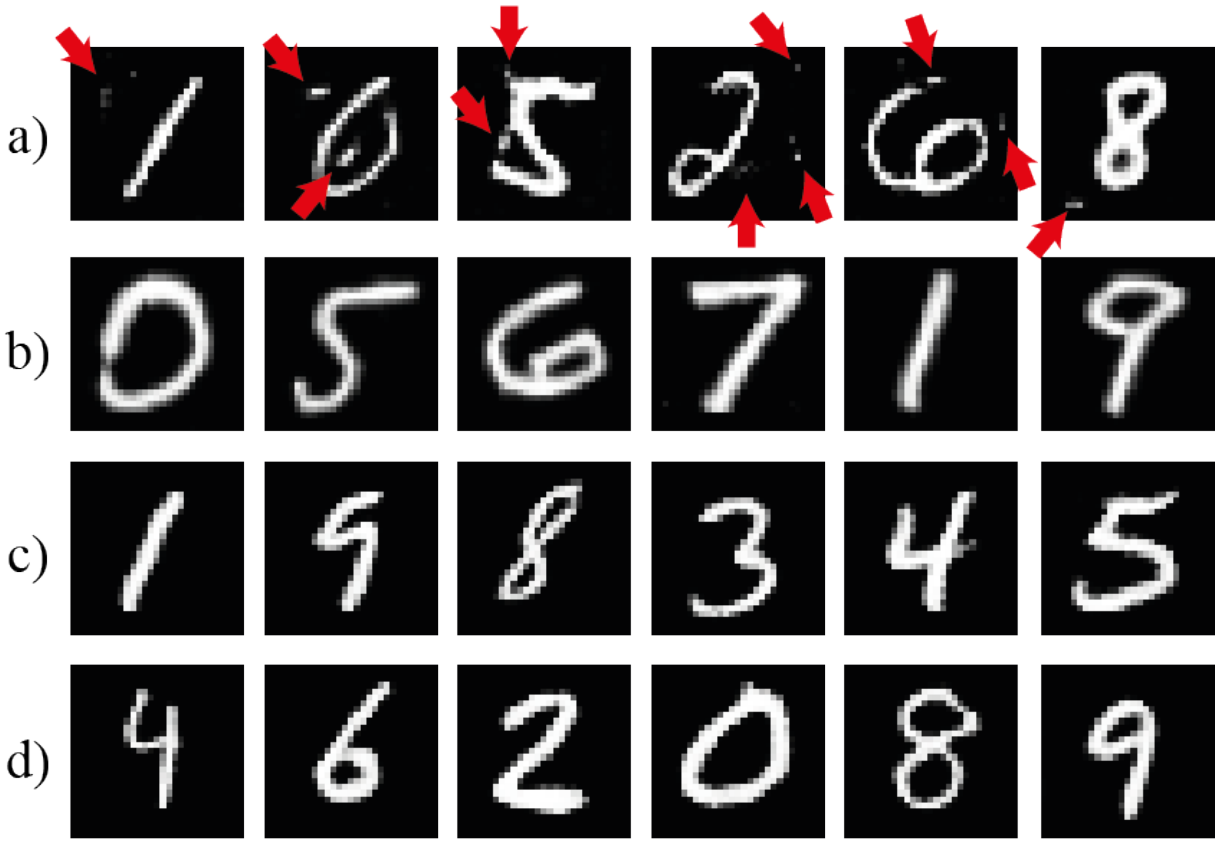


Figure 18 – Examples of artificial images generated by WGAN-GP (a), and by XWGAN-GP with the Saliency (b), DeepLIFT (c) and Gradient⊙Input (d) methods.

5.2 RESULTS ON HISTOPATHOLOGICAL DATASETS

We used XAI explanations to improve the quality of artificial images generated by GANs on histopathological datasets. Figure 19 to Figure 22 show some examples of original images and those generated by the GAN and XGAN models for the CR, LA, LG, and UCSB datasets, respectively. We evaluated the CR, LA, LG, and UCSB datasets using the FID and IS metrics. Table 2 shows the results obtained with the DCGAN, RAGAN, and WGAN-GP architectures. The results are organized by dataset, and the best results are highlighted in bold. 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. The green arrow indicates an improvement, while the red arrow indicates a decrease in the metric.

Considering the CR dataset (Figure 19), it is possible to note in Table 2 that the 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 FID and IS compared with RAGAN. The XRAGAN + DeepLIFT was the highlight, providing the lowest FID, 46.02, representing a

Table 2 – FID and IS scores for CR, LA, LG, and UCSB datasets.

		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** ↑

32.70% decrease compared to RAGAN (68.39). The XWGAN-GP also improved FID in all cases. The combination XWGAN-GP + Saliency provided the lowest FID, 72.01, 13.04% lower than WGAN-GP (82.81).

In the LA dataset (Figure 20), XDCGAN could not improve 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 were able to improve 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 could improve 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 21) and the performance using DCGAN-based architectures (Table 2), 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 FID and IS. The 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 22), 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 14.33% FID improvement. Considering the RAGAN-based models, XRAGAN + Saliency and XRAGAN + Gradient⊙Input improved FID and IS. The 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.

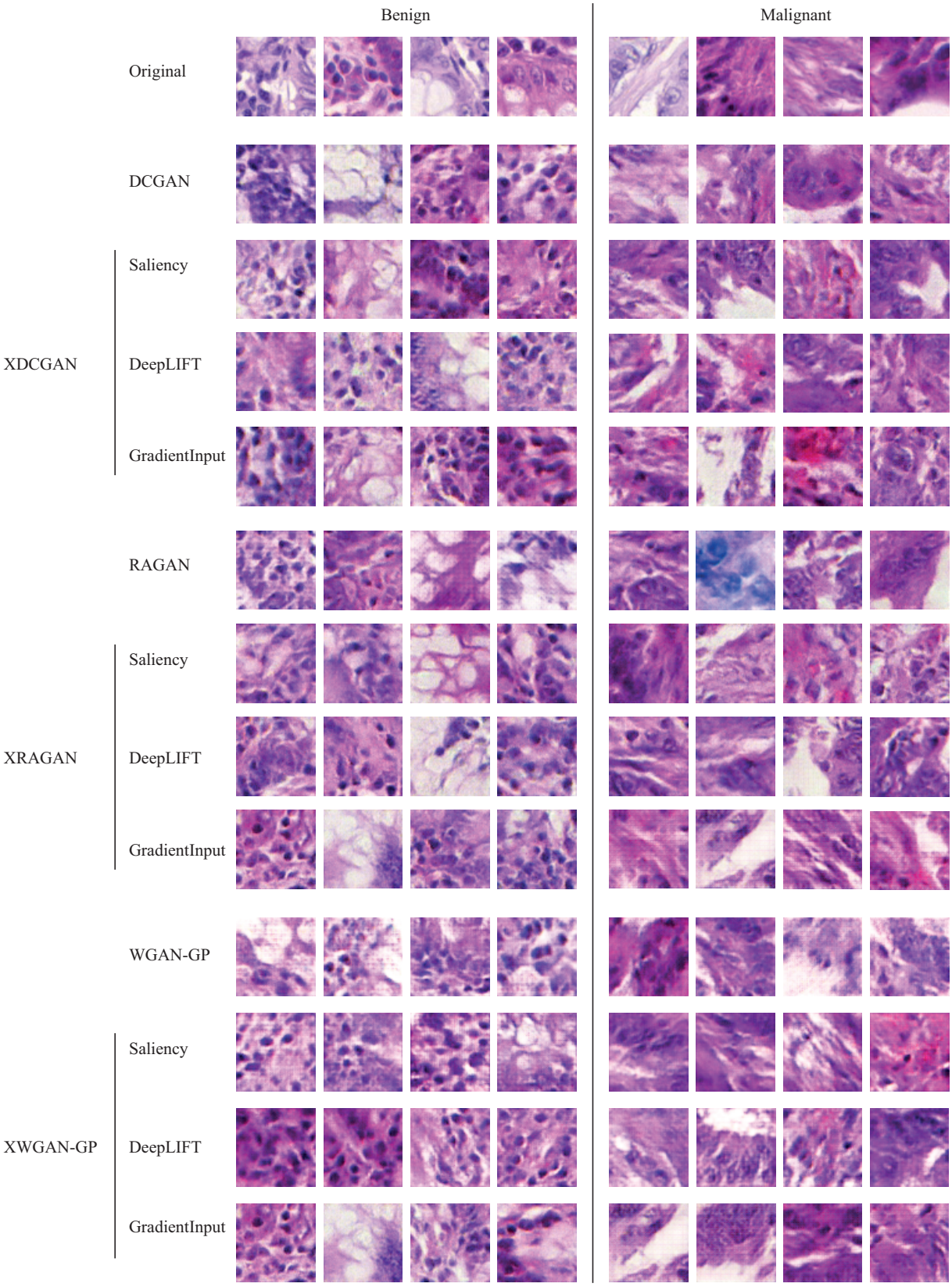


Figure 19 – Examples of generated images for the CR dataset.

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

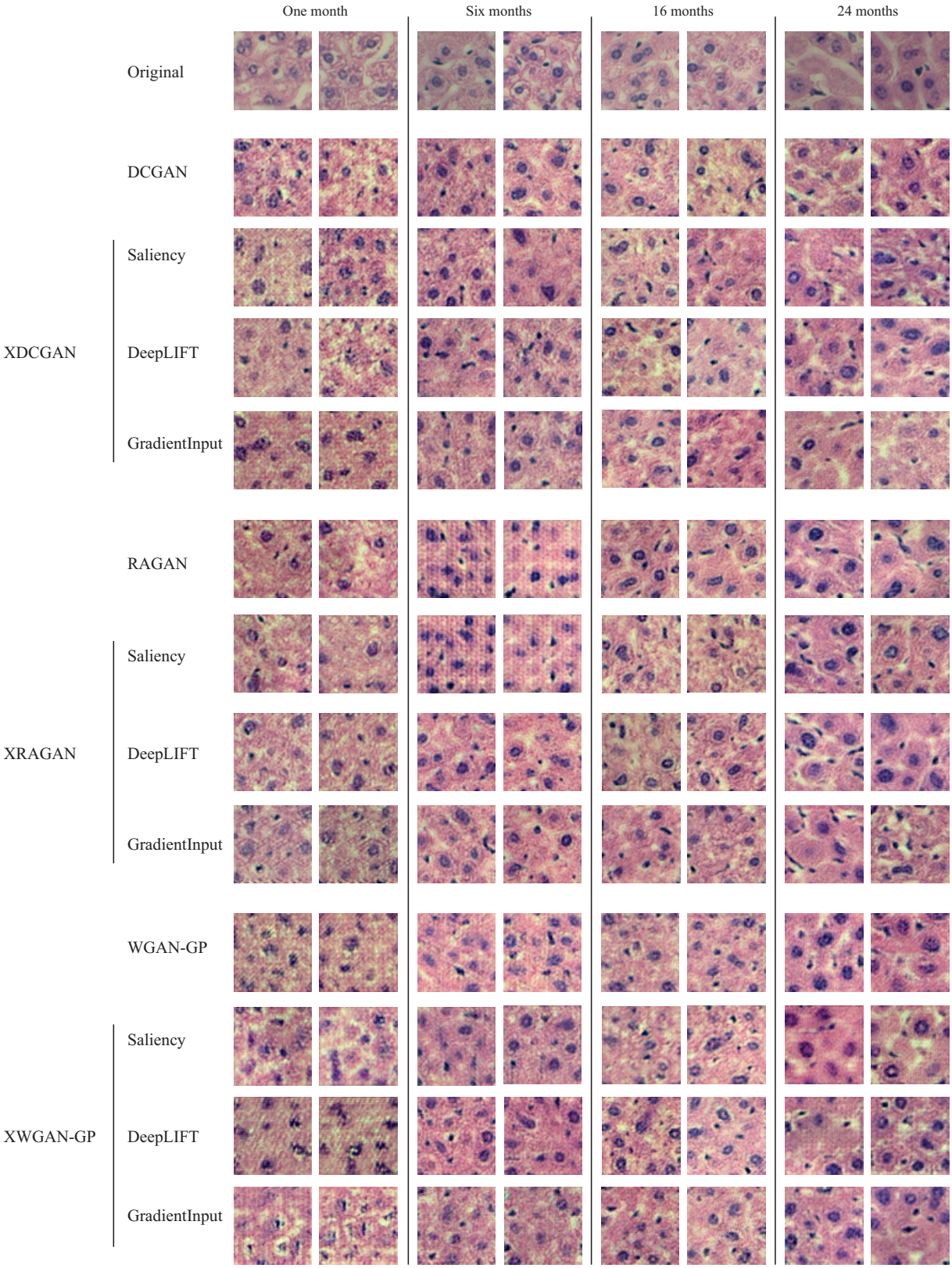


Figure 20 – Examples of generated images for the LA dataset.

models' classification performance. We followed the strategy described in section 4.6. Tables 3 to 6 show the classification results on the CR, LA, LG, and UCSB datasets, respectively. The table's first row presents the classification accuracy without data augmentation (without DA), and

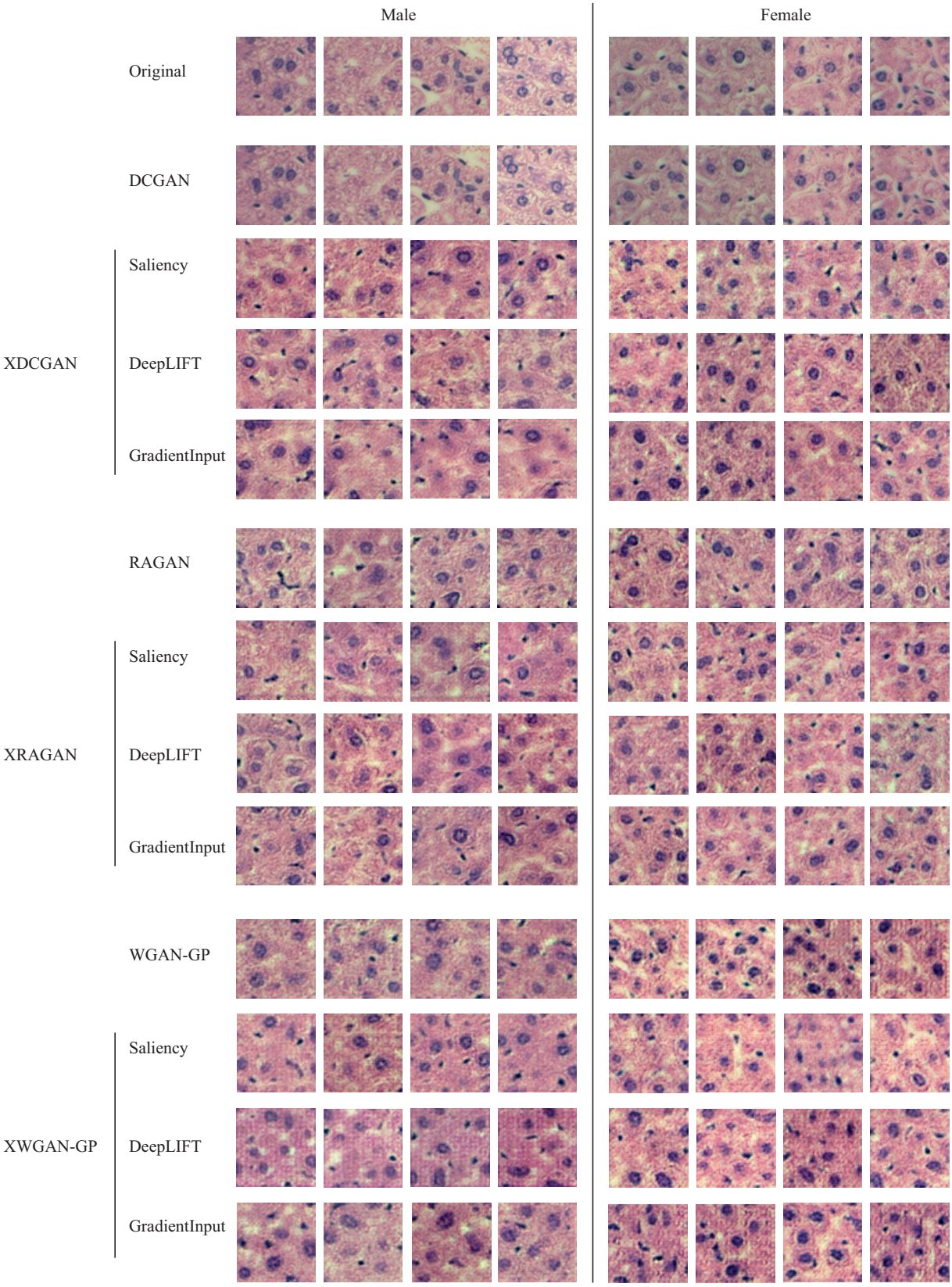


Figure 21 – Examples of generated images for the LG dataset.

results using GAN in data augmentation are organized by base architecture: DCGAN, RAGAN, and WGAN-GP. We considered the classification performance without data augmentation and

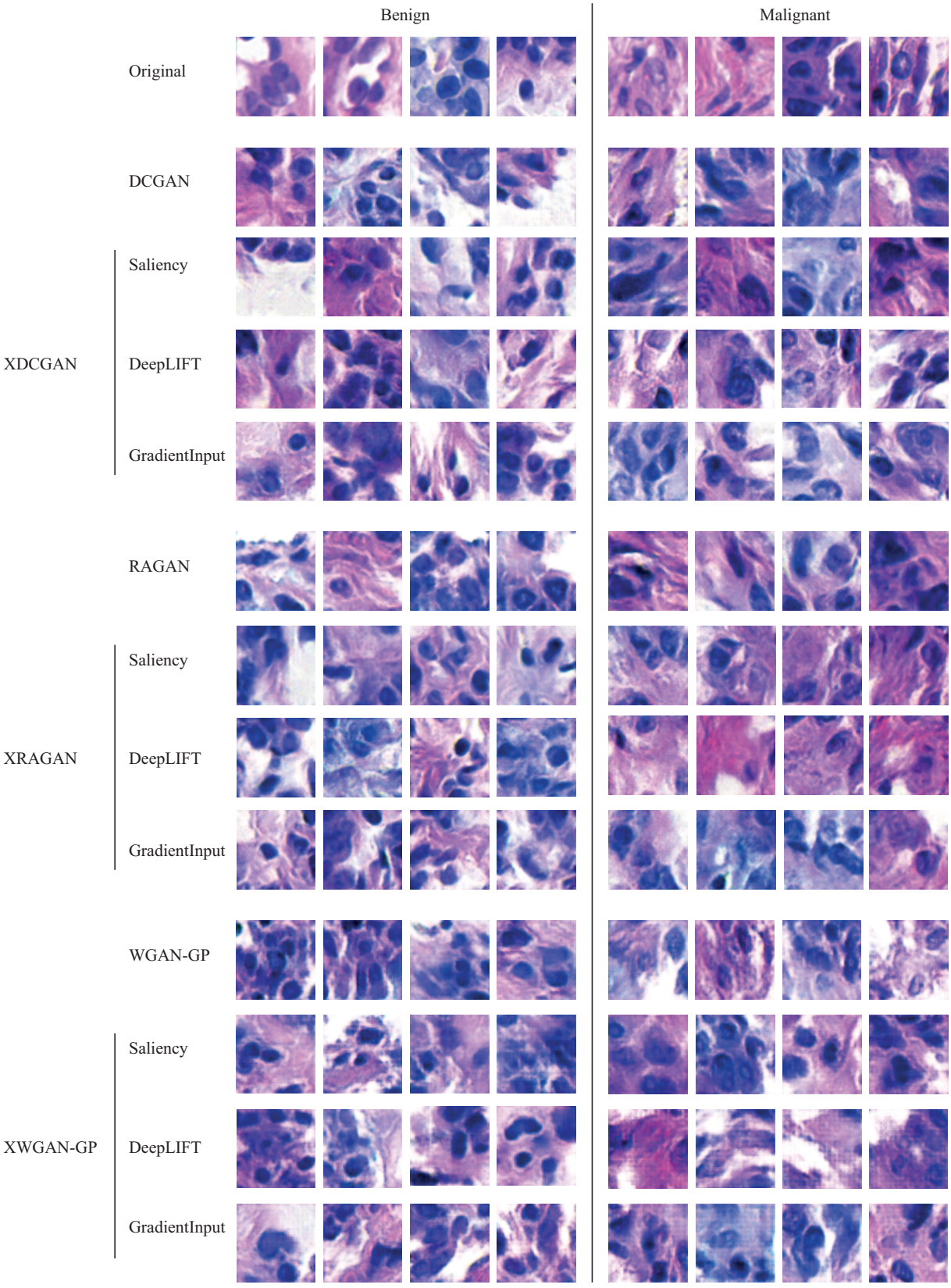


Figure 22 – Examples of generated images for the UCSB dataset.

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 Transformers-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.

Considering the classification results regarding the CR dataset (Table 3), 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 of 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%).

Table 3 – 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%

Regarding the LA dataset, XDCGAN could not improve FID (Table 2). The XDCGAN + Saliency, however, achieved the second-lowest FID (128.07) and and it is worth noting in Table 4 that this combination provided the best accuracy with DEiT (95.45%). It is also worth noting that the no GAN architecture 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 could improve DEiT, PVT, and CoAtNet classification performance compared to WGAN-GP. For instance, the 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 the WGAN-GP. Also, the combination XWGAN-GP + DeepLIFT, which provided the second-best FID and IS, achieved the best accuracy with CoAtNet, 99.34%.

Table 4 – 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 5 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 2), 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 2). 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 6 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 2). The 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 5 – 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 6 – 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%**

Lastly, 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 7 shows a ranking of the best combinations. Considering all datasets, the combination XWGAN-GP + Saliency outperformed WGAN-GP in seven cases. Second and third place 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 3 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 the Saliency, followed by Gradient$\odot$Input and DeepLIFT. These findings can guide researchers and experts in using GANs and XAI to develop artificial augmentation techniques. Our approach opens up new possibilities for enhancing data augmentation techniques and improving the overall performance of Transformer-based models in histopathological datasets.

Table 7 – Rank of the combinations with 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$\odot$Input	5
5	XDCGAN + Gradient$\odot$Input	4
6	XRAGAN + Gradient$\odot$Input	3
6	XRAGAN + DeepLIFT	3
7	XWGAN + DeepLIFT	1
7	XDCGAN + DeepLIFT	1

6 CONCLUSIONS

In this work, we proposed a new approach for training GANs using XAI-based methods to improve generation quality and data augmentation performance in histopathological datasets. We used gradient-based 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 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 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 than traditional GAN models. For instance, it was possible to decrease FID by up to 32.70% compared to the traditional architectures. This gain in quality positively affected the classification performance of Transformers models. It was possible to increase accuracy 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 the gradient-based methods' potential to improve the quality of image generation by GAN in histopathological datasets. We demonstrate that the features provided by these methods 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 are insights and guidelines for researchers and experts interested in developing artificial augmentation techniques for histopathological datasets.

In future work, we intend to conduct new tests using different combinations of GAN models and other ways to extract information from the images. Improving generator training is a subject that is still being explored, and there is much room for improvement. We also intend to explore other evaluation metrics, which can provide more insights into the quality of the generated images and the performance of the GAN models. We believe that the proposed method can be applied to other types of medical datasets or other contexts.

6.1 RESEARCH CONTRIBUTIONS

The development of this doctoral work was mainly carried out at the Department of Computing and Statistics of the São Paulo State University (Unesp), São José do Rio Preto campus,

under the supervision of Prof. Dr. Leandro Alves Neves. The student worked in collaboration with researchers from the University of Uberlândia (UFU), Brazil, and Federal University of São Paulo (UNIFESP), São José dos Campos, Brazil.

The student undertook a research internship for one year at the University of Bologna, Italy, where he collaborated with Prof. Dr. Alessandra Lumini. During his stay in Italy, the student participated in various research activities, such as attending scientific events. The student published and presented six papers at three international conferences, where he participated as a speaker and attendee:

- *Weeds Classification with Deep Learning: An Investigation Using CNN, Vision Transformers, Pyramid Vision Transformers, and Ensemble Strategy* (Rozendo et al., 2023), in collaboration with Prof. Dr. Alessandra Lumini, from Università di Bologna, Italy. It was presented at the event CIARP 2023, held in Porto, Portugal in November 2023.

- *CNN Ensembles for Nuclei Segmentation on Histological Images of OED* (Silva et al., 2023), which was presented at the event CBMS 2023, held in L'Aquila, Italy in June 2023.

- *Handcrafted features vs deep-learned features: Hermite Polynomial Classification of Liver Images* (Pereira et al., 2023), which was presented at the event CBMS 2023, held in L'Aquila, Italy in June 2023.

- *An Investigation of Deep-Learned Features for Classifying Radiographic Images of COVID-19* (Miguel et al., 2023), which was presented at the event ICEIS 2023, held in Praga, Czech Republic in April 2023.

- *Classification of H&E images via CNN models with XAI approaches, deepdream representations and multiple classifiers* (Neves et al., 2023), which was presented at the event ICEIS 2023, held in Praga, Czech Republic in April 2023.

- *X-GAN: Generative Adversarial Networks Training Guided with Explainable Artificial Intelligence* (Rozendo. et al., 2024), which was presented at the event ICEIS 2024, held in Angers, France in April 2024.

He also published two papers in international journals:

- *Data Augmentation in Histopathological Classification: An Analysis Exploring GANs with XAI and Vision Transformers* in the journal Applied Sciences (Rozendo et al., 2024), in collaboration with Prof. Dr. Alessandra Lumini, from Università di Bologna, Italy.

- *Exploring DeepDream and XAI Representations for Classifying Histological Images* (Martinez et al., 2024) in the journal SN Computer Science.

Finally, the student presented a poster at the XII Workshop do Programa de Pós-Graduação em Ciência da Computação da Unesp (WPPGCC 2023), held on November 2023 remotely. The poster was entitled *X-GAN: Generative Adversarial Networks Training Guided with Explainable Artificial Intelligence*.

BIBLIOGRAPHY

AHMAD, B. et al. Brain tumor classification using a combination of variational autoencoders and generative adversarial networks. **Biomedicines**, Multidisciplinary Digital Publishing Institute, v. 10, n. 2, p. 223, 2022.

ANWAR, S. M. et al. Medical image analysis using convolutional neural networks: a review. **Journal of medical systems**, Springer, v. 42, p. 1–13, 2018.

APOSTOLOPOULOS, I. D.; PAPATHANASIOU, N. D.; PANAYIOTAKIS, G. S. Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning. **Biocybernetics and Biomedical Engineering**, Elsevier, v. 41, n. 4, p. 1243–1257, 2021.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. Wasserstein generative adversarial networks. In: PMLR. **International conference on machine learning**. [S.l.], 2017. p. 214–223.

ATAKY, S. T. M. et al. Data augmentation for histopathological images based on gaussian-laplacian pyramid blending. In: **2020 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2020. p. 1–8.

BAI, Q. et al. Glead: Improving gans with a generator-leading task. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2023. p. 12094–12104.

Barredo Arrieta, A. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. **Information Fusion**, v. 58, p. 82–115, 2020. ISSN 1566-2535.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 8, p. 1798–1828, 2013.

BORJI, A. Pros and cons of gan evaluation measures. **Computer Vision and Image Understanding**, v. 179, p. 41–65, 2019. ISSN 1077-3142. Disponível em: https://www.sciencedirect.com/science/article/pii/S1077314218304272.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. **Electronics**, v. 8, n. 8, 2019. ISSN 2079-9292. Disponível em: https://www.mdpi.com/2079-9292/8/8/832.

ÇELIK, G.; TALU, M. F. A new 3d mri segmentation method based on generative adversarial network and atrous convolution. **Biomedical Signal Processing and Control**, Elsevier, v. 71, p. 103155, 2022.

CHEN, X. et al. Infogan: interpretable representation learning by information maximizing generative adversarial nets. In: **Proceedings of the 30th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2016. (NIPS'16), p. 2180–2188. ISBN 9781510838819.

CRESWELL, A. et al. Generative adversarial networks: An overview. **IEEE Signal Processing Magazine**, IEEE, v. 35, n. 1, p. 53–65, 2018.

DAI, Z. et al. Coatnet: Marrying convolution and attention for all data sizes. In: RANZATO, M. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2021. v. 34, p. 3965–3977.

DARGAN, S. et al. A survey of deep learning and its applications: a new paradigm to machine learning. **Archives of Computational Methods in Engineering**, Springer, v. 27, n. 4, p. 1071–1092, 2020.

DING, H.; HUANG, N.; CUI, X. Leveraging gans data augmentation for imbalanced medical image classification. **Applied Soft Computing**, v. 165, p. 112050, 2024. ISSN 1568-4946.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: **International Conference on Learning Representations**. [s.n.], 2021. Disponível em: https://openreview.net/forum?id=YicbFdNTTy.

FRID-ADAR, M. et al. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. **Neurocomputing**, Elsevier, v. 321, p. 321–331, 2018.

FUKUSHIMA, K. Cognitron: A self-organizing multilayered neural network. **Biological cybernetics**, Springer, v. 20, n. 3, p. 121–136, 1975.

GARCEA, F. et al. Data augmentation for medical imaging: A systematic literature review. **Computers in Biology and Medicine**, v. 152, p. 106391, 2023. ISSN 0010-4825.

GELASCA, E. D. et al. Evaluation and benchmark for biological image segmentation. In: **2008 15th IEEE International Conference on Image Processing**. [S.l.: s.n.], 2008. p. 1816–1819.

GONZALEZ, R.; WOODS, R. **Digital Image Processing Global Edition**. Pearson Deutschland, 2017. 1024 p. ISBN 9781292223049. Disponível em: https://elibrary.pearson.de/book/99.150005/9781292223070.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. ISBN 9780262035613. Disponível em: http://www.deeplearningbook.org.

GOODFELLOW, I. et al. Generative adversarial nets. In: GHAHRAMANI, Z. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2014. v. 27. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

GUERRERO, R. E. D. et al. A data augmentation methodology to reduce the class imbalance in histopathology images. **Journal of Imaging Informatics in Medicine**, Springer, p. 1–16, 2024.

GULRAJANI, I. et al. Improved training of wasserstein gans. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

GURCAN, M. N. et al. Histopathological image analysis: A review. **IEEE Reviews in Biomedical Engineering**, v. 2, p. 147–171, 2009.

HALL, J. E.; HALL, M. E. **Guyton and Hall textbook of medical physiology e-Book**. [S.l.]: Elsevier Health Sciences, 2020.

HE, K. et al. Transformers in medical image analysis. **Intelligent Medicine**, v. 3, n. 1, p. 59–78, 2023. ISSN 2667-1026. Disponível em: https://www.sciencedirect.com/science/article/pii/S2667102622000717.

HE, L. et al. Histology image analysis for carcinoma detection and grading. **Computer Methods and Programs in Biomedicine**, v. 107, n. 3, p. 538–556, 2012. ISSN 0169-2607. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169260711003245.

HEUSEL, M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v. 30.

HUMEAU-HEURTIER, A. Texture feature extraction methods: A survey. **IEEE Access**, v. 7, p. 8975–9000, 2019.

JABBAR, A.; LI, X.; OMAR, B. A survey on generative adversarial networks: Variants, applications, and training. **ACM Computing Surveys (CSUR)**, ACM New York, NY, v. 54, n. 8, p. 1–49, 2021.

JAIN, A. K.; MAO, J.; MOHIUDDIN, K. M. Artificial neural networks: A tutorial. **Computer**, IEEE, v. 29, n. 3, p. 31–44, 1996.

JOLICOEUR-MARTINEAU, A. The relativistic discriminator: a key element missing from standard GAN. In: **International Conference on Learning Representations**. [s.n.], 2019. Disponível em: https://openreview.net/forum?id=S1erHoR5t7.

JOTHI, J. A. A.; RAJAM, V. M. A. A survey on automated cancer diagnosis from histopathology images. **Artificial Intelligence Review**, Springer, v. 48, p. 31–81, 2017.

KEBAILI, A.; LAPUYADE-LAHORGUE, J.; RUAN, S. Deep learning approaches for data augmentation in medical imaging: a review. **Journal of Imaging**, MDPI, v. 9, n. 4, p. 81, 2023.

KRIZHEVSKY, A. **Learning Multiple Layers of Features from Tiny Images**. 2009. Disponível em: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 60, n. 6, p. 84–90, maio 2017. ISSN 0001-0782.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LI, Z. et al. A survey of convolutional neural networks: Analysis, applications, and prospects. **IEEE Transactions on Neural Networks and Learning Systems**, v. 33, n. 12, p. 6999–7019, 2022.

LIAO, Y.-T. et al. Data augmentation based on generative adversarial networks to improve stage classification of chronic kidney disease. **Applied Sciences**, Multidisciplinary Digital Publishing Institute, v. 12, n. 1, p. 352, 2022.

LIU, W. et al. A survey of deep neural network architectures and their applications. **Neurocomputing**, v. 234, p. 11–26, 2017. ISSN 0925-2312.

LORENCIN, I. et al. On urinary bladder cancer diagnosis: Utilization of deep convolutional generative adversarial networks for data augmentation. **Biology**, Multidisciplinary Digital Publishing Institute, v. 10, n. 3, p. 175, 2021.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v. 30.

MADANI, A. et al. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In: IEEE. **2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)**. [S.l.], 2018. p. 1038–1042.

MAHENDRAN, A.; VEDALDI, A. Understanding deep image representations by inverting them. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 5188–5196.

MARTINEZ, J. M. C. et al. Exploring deepdream and xai representations for classifying histological images. **SN Computer Science**, Springer, v. 5, n. 4, p. 362, 2024.

MIGUEL, P. L. et al. An investigation of deep-learned features for classifying radiographic images of covid-19. In: **ICEIS**. [S.l.: s.n.], 2023. p. 675–682.

MINH, D. et al. Explainable artificial intelligence: a comprehensive review. **Artificial Intelligence Review**, Springer, p. 1–66, 2022.

MORDVINTSEV, A.; OLAH, C.; TYKA, M. **Inceptionism: Going Deeper into Neural Networks**. 2015. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Accessed: 13-02-2022.

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**, v. 27, p. 100779, 2021. ISSN 2352-9148.

MUMUNI, A.; MUMUNI, F. Data augmentation: A comprehensive survey of modern approaches. **Array**, Elsevier, v. 16, p. 100258, 2022.

NASH, J. F. Equilibrium points in n-person games. **Proceedings of the National Academy of Sciences**, v. 36, n. 1, p. 48–49, 1950. Disponível em: https://www.pnas.org/doi/abs/10.1073/pnas.36.1.48.

NEVES, L. A. et al. Classification of he images via cnn models with xai approaches, deepdream representations and multiple classifiers. In: INSTICC. **Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS**. [S.l.]: SciTePress, 2023. p. 354–364. ISBN 978-989-758-648-4. ISSN 2184-4992.

NGUYEN, A. et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. **Advances in neural information processing systems**, v. 29, 2016.

NIELSEN, I. E. et al. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. **IEEE Signal Processing Magazine**, v. 39, n. 4, p. 73–84, 2022.

ODENA, A.; OLAH, C.; SHLENS, J. Conditional image synthesis with auxiliary classifier GANs. In: PRECUP, D.; TEH, Y. W. (Ed.). **Proceedings of the 34th International Conference on Machine Learning**. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 2642–2651. Disponível em: https://proceedings.mlr.press/v70/odena17a.html.

OLAH, C.; MORDVINTSEV, A.; SCHUBERT, L. Feature visualization. **Distill**, 2017. Https://distill.pub/2017/feature-visualization.

OLAH, C. et al. The building blocks of interpretability. **Distill**, 2018. Https://distill.pub/2018/building-blocks.

PARVAIZ, A. et al. Vision transformers in medical computer vision—a contemplative retrospection. **Engineering Applications of Artificial Intelligence**, v. 122, p. 106126, 2023. ISSN 0952-1976. Disponível em: https://www.sciencedirect.com/science/article/pii/S095219762300310X.

PEREIRA, D. C. et al. Handcrafted features vs deep-learned features: Hermite polynomial classification of liver images. In: **2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)**. [S.l.: s.n.], 2023. p. 495–500.

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: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322.

ROZENDO., G. et al. X-gan: Generative adversarial networks training guided with explainable artificial intelligence. In: INSTICC. **Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS**. [S.l.]: SciTePress, 2024. p. 674–681. ISBN 978-989-758-692-7. ISSN 2184-4992.

ROZENDO, G. B. et al. Data augmentation in histopathological classification: An analysis exploring gans with xai and vision transformers. **Applied Sciences**, v. 14, n. 18, 2024. ISSN 2076-3417. Disponível em: https://www.mdpi.com/2076-3417/14/18/8125.

ROZENDO, G. B. et al. Weeds classification with deep learning: An investigation using cnn, vision transformers, pyramid vision transformers, and ensemble strategy. In: SPRINGER. **Iberoamerican Congress on Pattern Recognition**. [S.l.], 2023. p. 229–243.

SALAU, A. O.; JAIN, S. Feature extraction: A survey of the types, techniques, applications. In: **2019 International Conference on Signal Processing and Communication (ICSC)**. [S.l.: s.n.], 2019. p. 158–164.

SALIMANS, T. et al. Improved techniques for training gans. In: LEE, D. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2016. v. 29.

SARVAMANGALA, D.; KULKARNI, R. V. Convolutional neural networks in medical image understanding: a survey. **Evolutionary intelligence**, Springer, v. 15, n. 1, p. 1–22, 2022.

SHAMIR, L. et al. Iicbu 2008: a proposed benchmark suite for biological image analysis. **Medical & biological engineering & computing**, Springer, v. 46, p. 943–947, 2008.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. In: PRECUP, D.; TEH, Y. W. (Ed.). **Proceedings of the 34th International Conference on Machine Learning**. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 3145–3153. Disponível em: https://proceedings.mlr.press/v70/shrikumar17a.html.

SILVA, A. B. et al. Cnn ensembles for nuclei segmentation on histological images of oed. In: **2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)**. [S.l.: s.n.], 2023. p. 601–604.

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)**. [S.l.]: ICLR, 2014. p. 1–8.

SIRINUKUNWATTANA, K. et al. Gland segmentation in colon histology images: The glas challenge contest. **Medical Image Analysis**, v. 35, p. 489–502, 2017. ISSN 1361-8415.

SOUZA, V. L. T. de et al. A review on generative adversarial networks for image generation. **Computers & Graphics**, v. 114, p. 13–25, 2023. ISSN 0097-8493.

SUN, Y. et al. Lesion segmentation in gastroscopic images using generative adversarial networks. **Journal of Digital Imaging**, Springer, p. 1–10, 2022.

TOUVRON, H. et al. Training data-efficient image transformers & distillation through attention. In: MEILA, M.; ZHANG, T. (Ed.). **Proceedings of the 38th International Conference on Machine Learning**. [S.l.]: PMLR, 2021. (Proceedings of Machine Learning Research, v. 139), p. 10347–10357.

WANG, J. et al. Improving gan equilibrium by raising spatial awareness. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2022. p. 11285–11293.

WANG, J. et al. A review of deep learning on medical image analysis. **Mobile Networks and Applications**, Springer, v. 26, n. 1, p. 351–380, 2021.

WANG, W. et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2021. p. 568–578.

WANG, Z.; SHE, Q.; WARD, T. E. Generative adversarial networks in computer vision: A survey and taxonomy. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 54, n. 2, feb 2021. ISSN 0360-0300.

YAMASHITA, R. et al. Convolutional neural networks: an overview and application in radiology. **Insights into imaging**, Springer, v. 9, n. 4, p. 611–629, 2018.

YI, X.; WALIA, E.; BABYN, P. Generative adversarial network in medical imaging: A review. **Medical Image Analysis**, v. 58, p. 101552, 2019. ISSN 1361-8415.

ZARE, M. R.; ALEBIOSU, D. O.; LEE, S. L. Comparison of handcrafted features and deep learning in classification of medical x-ray images. In: **2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)**. [S.l.: s.n.], 2018. p. 1–5.

ZHU, J.-Y. et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 2223–2232.