NOVA
IMS
Information
Management
School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Data Science and Advanced Analytics

Impact of GAN-based Lesion-Focused Medical Image Super-Resolution on Radiomic Feature Robustness

Erick Costa de Farias

Dissertation submitted in partial fulfillment of the requirements for the degree of

Master of Science in Data Science and Advanced Analytics with specialization in Data Science

Supervisor: Prof. Dr. Mauro Castelli

NOVA Information Management School
Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

Impact of GAN-based Lesion-Focused Medical Image Super-Resolution on Radiomic Feature Robustness

Acknowledgements

I am immensely grateful for Professor Mauro Castelli supervision. This project was developed in the middle of a pandemic and wouldn't be concluded without his insightful, supportive and empathetic guidance, which prioritized my personal well being at all times while I struggled with technical and emotional challenges.

I am also deeply indebted to Leonardo Rundo and Changhee Han, brilliant researchers and advisors who made this journey incredible with their perfect blend of insight and humor. I also extend my gratitude to Christian Di Noia, who contributed to this research with his expertise in radiomics analysis.

I'd be remiss not to mention Dad and Mom, who made unmeasureable sacrifices so I could have the opportunities that led me to this moment. For them I make my daily existence an act of gratitude.

And lastly, I thank God, in whom I find my passion to seek wisdom, knowledge and love for my fellows.

This work was partially supported by The Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre [C9685/A25177], the Wellcome Trust Innovator Award, UK [215733/Z/19/Z] and the CRUK National Cancer Imaging Translational Accelerator (NCITA) [C42780/A27066]. Additional support was also provided by the National Institute of Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014). This works was also finantially supported by national funds through the FCT (Fundação para a Ciência e a Tecnologia) by the projects GADgET (DSAIPA/DS/0022/2018) and the Slovenian Research Agency (research core funding no. P5-0410).

“The most incomprehensible fact is the fact that we comprehend at all.” (Abraham Joshua Heschel)

Abstract

Robust machine learning models based on radiomic features might allow for accurate diagnosis, prognosis, and medical decision-making. Unfortunately, the lack of standardized radiomic feature extraction has hampered their clinical use. Since the radiomic features tend to be affected by low voxel statistics in regions of interest, increasing the sample size would improve their robustness in clinical studies. Therefore, we propose a Generative Adversarial Network (GAN)-based lesion-focused framework for Computed Tomography (CT) image Super-Resolution (SR); for the lesion (i.e., cancer) patch-focused training, we incorporate Spatial Pyramid Pooling (SPP) into GAN-Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE). At 2× SR, the proposed model achieved better perceptual quality with less blurring than the other considered state-of-the-art SR methods, while producing comparable results at 4× SR. We also evaluated the robustness of our model's radiomic feature in terms of quantization on a different lung cancer CT dataset using Principal Component Analysis (PCA). Intriguingly, the most important radiomic features in our PCA-based analysis were the most robust features extracted on the GAN-super-resolved images. These achievements pave the way for the application of GAN-based image Super-Resolution techniques for studies of radiomics for robust biomarker discovery.

Keywords: radiomics, super-resolution, gan, deep-learning

Resumo

Modelos de machine learning robustos baseados em atributos radiômicos possibilitam diagnósticos e decisões médicas mais precisas. Infelizmente, por causa da falta de padronização na extração de atributos radiômicos, sua utilização em contextos clínicos tem sido restrita. Considerando que atributos radiômics tendem a ser afetados pelas estatítiscas de voxels de baixo volume nas regiões de interesse, o aumento to tamanho da amostra tem o potencial de melhorar a robustez desses atributos em estudos clínicos. Portanto, esse trabalho propões um framework baseado numa rede neural generativa (GAN) focada na região de interesse para a super-resolução de imagens de Tomografia Computadorizada (CT). Para treinar a rede de forma concentrada na lesão (i.e. cancer), incorporamos a tecnica de Spatial Pyramid Pooling no framework da GAN-CIRCLE. Nos experimentos de super-resolução 2×, o modelo proposto alcançou melhor qualidade perceptual com menos embaçamento do que outros métodos estado-da-arte considerados. A robustez dos atributos radiômics das imagens super-resolvidas geradas pelo modelo também foram analizadas em termos de quantização em um banco de imagens diferente, contendo imagens de tomografia computadorizada de câncer de pulmão, usando anaálise de componentes principaiss (PCA). Intrigantemente, os atributos radiômicos mais importantes nessa análise foram também os atributos mais robustos extraídos das imagens super-resolvidas pelo método proposto. Esses resultados abrem caminho para a aplicação de técnicas de super-resolução baseadas em redes neurais generativas aplicadas a estudos de radômica para a descoberta de biomarcadores robustos.

Palavras-chave: radiômica, super-resolução, redes generativas, apredizado de máquina, apredizagem profunda,

Contents

Chapter

1 Introduction

Recent years have witnessed an increasing interest in the area of medical image analysis. This growth was driven by the availability of datasets with a vast number of images and by the development of appropriate (statistical and artificial intelligence-based) tools for analyzing these large-scale datasets.

In particular, instead of focusing on the visual interpretation of an image, it is nowadays common to extract quantitative features from the images and, subsequently, to apply machine learning techniques for obtaining meaningful insights into the clinical problem at hand. This process is known as radiomics(1). Robust machine learning models based on large-scale radiomic features might allow for accurate diagnosis, prognosis, and medical decision-making; of course, thoroughly considering the whole radiomic processes is essential to obtain these reliable models.

Despite the potential of radiomics, high quantitative feature variability across different software implementations has hampered its clinical use(2, 3).

This phenomenon derives from the lack of standardized definitions and extraction of radiomic features with validated reference values. To tackle this limitation and facilitate clinical interpretation, the Image Biomarker Standardization Initiative(2) produced and validated the reference values for commonly-used radiomic features. However, as the paper's authors highlighted, image features still need to be robust against differences in acquisition, reconstruction, and segmentation to ensure reproducibility. For this reason, recent studies have investigated the robustness of radiomic features in several scenarios and applications using heterogeneous datasets. Several sources of variability have been assessed, such as image and region of interest (ROI) perturbations(4, 5), slice thickness variations(6, 7), and different resampling strategies(8). Since the radiomic features might tend to be affected by low statistics in ROI voxels, we hypothesize that increasing such a sample size would increase the robustness of radiomic features in clinical studies. Therefore, we aim to apply image Super-Resolution

(SR) to increase the number of voxels used in the computation of radiomic features.

The availability of a vast amount of images may contribute to achieving reference values for radiomic features, thus improving their robustness. Unfortunately, in some clinical settings, labeled data is scarce and expensive to create. To overcome this limitation, recent contributions proposed the use of data augmentation techniques based on GANs. Sandfort *et al.*(9) used CycleGAN(10)-based DA for Computed Tomography (CT) segmentation by translating contrast images into synthetic non-contrast ones. Experimental results showed that, in several segmentation tasks, performance improved significantly. Thus, CycleGANs represent a viable method to reduce manual segmentation effort and cost in computed tomography imaging and improve feature robustness.

To maximize the data augmentation effect with the GAN combinations, Han and coauthors(11) proposed a two-step GAN-based data augmentation that generates and refines brain magnetic resonance images with/without tumors separately. To assess the performance of their method, authors investigated convolutional neural network (CNN)-based tumor classification results. Experimental results show that, when combined with classic data augmentation, the proposed two-step GAN-based data augmentation technique outperforms the classic data augmentation alone in tumor detection and other medical imaging tasks.

The most prominent work on CT image SR is GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE)(12), outperforming previous works(13–16). GAN-CIRCLE can preserve anatomical information and suppress noise, leading to excellent diagnostic performance in terms of traditional image quality metrics(12, 17).

For example, Guha *et al.*(17) exploited GAN-CIRCLE to super-resolve trabecular bone microstructures and improved the structural similarity index. Meanwhile, GAN-based lesion-focused medical image SR can improve SR performance around lesions, especially for downstream radiomic analyses(18). Along with GAN-based medical image SR, novel approaches based on progressive GANs(19) and attention mechanisms(20) have been recently applied to video SR.

For the first time, in this paper, we evaluate the robustness of radiomic features extracted from super-resolved images by GAN-SR and bicubic interpolation. The authors incorporated Spatial Pyramid Pooling (SPP)(21) into the discriminator of GAN-CIRCLE(12) to handle different input CT image sizes for patch-focused training in lesions; we cropped the input CT images to their lesion bounding boxes to reduce training costs and improve image quality (e.g., fewer artifacts)(18). Along with perceptual quality evaluation, we also assessed the robustness of radiomics, in terms of quantization, for our model against a bicubic interpolation baseline on a separate lung cancer CT dataset. We found that the most important radiomic features in our Principal Component Analysis (PCA)-based examination were the most robust features extracted on the GAN-super-resolved images.

To summarize, this work provides the following contributions:

- definition of the first GAN-based, lesion-focused, SR framework for CT images;
- comparison with state-of-the-art SR techniques highlighting the suitability of the proposed framework;
- at 2× SR, the images are characterized by better perceptual quality, as suggested by the peak signal-to-noise ratio and structural similarity index measures, on a large-scale dataset;
- at 4× SR, the proposed GAN-based model achieves comparable results to the ones obtained by state-of-the-art SR techniques;
- the proposed GAN-SR framework improves the robustness of the most important radiomic features in an independent lung CT dataset.

Chapter

2 Theoretical Framework

2.1 Image super-resolution

Digital images are composed of pixels, and the term image spatial resolution refers to the number of pixels per unit distance. Higher resolution images contain more detail than lower resolution images.
High-resolution images play a critical role in many areas, like astronomic image processing, microscopic image processing, medical image processing, and the media industry. However, obtaining these high-resolution images frequently depends on better-quality image sensors at a high cost. Obtaining a high-resolution (HR) image from a low-resolution (LR) image through the signal processing and computational enhancement of the resolution in order to overcome these limitations is the main objective of the image super-resolution (SR) task (*22*, *23*).

One of the main challenges of the SR task is the reconstruction of the actual HR image represented by the LR counterpart, as the SR problem is ill-posed: there are virtually an infinite number of HR images that can be reconstructed from a given LR image. The ill-posedness of the SR problem is caused by the aliasing effect of the down-sampling process(*22*). Aliasing occurs when the sampling rate of a signal is lower than twice its highest frequency component, causing a false representation of the signal.

Given a high-resolution image I_{HR}, its low-resolution counterpart I_{LR} is generally defined through the relationship $f(I_{HR}) = I_{LR}$, where the mapping $f : I_{HR} \to I_{LR}$ is an unknown degradation process composed of geometric transformations, blurring, down-sampling and noise addition. The single image super-resolution (SISR) problem has been conventionally formulated as follows:

$$x = D \times B \times M \times y + n \tag{2.1}$$

where x denotes the HR image, y denotes the LR image, D is the down-sampling filter, B is the blurring filter, M is the warping filter, and n n represents additive white Gaussian noise (AWGN). Our goal is to learn the mapping $g : I_{LR} \rightarrow I_{HR}$, which is the approximation of f^{-1} (24)).

The super-resolution task is essential for medical image processing, as medical images often contain a high pixel density, which can offer more detail and be critical for applications in medical imaging. The improvement of such images is pivotal to many diagnostic and prognostic processes, as the enhanced images are expected to unveil important information that would not be identifiable in the raw image. For instance, Computerized Tomography (CT) scan images have their image resolution restrained by the scanning technology, which makes it challenging to observe the details of the images, and even with advanced visualization software, the resolution is still lower than the ideal for crucial tasks such as early tumor detection (12). Obtaining CT images with better resolution is associated with high hardware costs and elevated radiation dose in patients, which could generate genetic damages and other diseases.

The SISR for medical images is still challenging since medical images have lower signal-to-noise ratios than natural images. Additionally, Deep Learning based models pre-trained on natural images may synthesize unrealistic patterns in medical images, which could affect the clinical interpretation and diagnosis.

SR techniques are conventionally divided into three broad categories: interpolation-based, learning-based, and model-based:

Interpolation-based methods Many researchers in the past have addressed the SISR problem, and many methods have been proposed to solve it. The first methods proposed to solve the SISR problem were based on interpolation. Interpolation approaches are based on interpolating the LR image to obtain an HR image with higher resolution (22, 25), such as bilinear interpolation, bicubic and nearest-neighbor interpolation (22). They are simple, fast, and frequently used as a baseline for many SR studies. However, these methods tend to generate exceedingly smooth images with rough artifacts, which worsens with larger scales. It is important to note, though, that interpolation methods cannot recover components lost or degraded during the downsampling process, and in some sources, they are not even considered SR techniques (22). The most common interpolation methods are the nearest-neighbor, bilinear, and bicubic(25):

Nearest-neighbor interpolation is a simple technique that generates HR images by repeating the pixels of the LR image. This technique is the most used to enlarge images in common digital image software because it does not change the color information. It is a fast approach that does not require any training, but it is also the most straightforward interpolation technique, and for this reason, it generates images with rough artifacts (22).

Bilinear interpolation is a weighted average of the values at the four corners of the rectangle. For an (x,y) position inside the rectangle, the weights are determined by the distance between the point and the corners. This method produces few artifacts, as it has an anti-aliasing effect.

Bicubic interpolation works by using a weighted average of the pixels in the 4x4 neighborhood surrounding the target pixel. The weighting is based on a cubic function, which gives more weight to the pixels closest to the target pixel and results in a smoother image with less aliasing than other methods, such as bilinear interpolation.

Model-based methods: Model-based methods are based on modeling the degradation process and finding a solution to a specific objective function that best describes the relationship between the HR and LR images. These objective functions are based on the natural image statistics, such as the image gradient, coherence, and sparsity.

Learning-based methods: In learning-based methods, the basic idea is to find a mapping between the low-resolution image and the high-resolution image. This mapping can be learned by training on a known LR-HR pair of images. Once the mapping is learned, the LR image can be enhanced by reconstructing it through this mapping. Jiang (26), for instance, proposed a dictionary learning and sparse representation approach, which presented strong robustness in preserving image features and outperformed interpolation methods.

With the advancement of machine learning, the usage of deep learning received attention in SR research, and many deep-learning architectures have been proposed to solve SR problems. The first deep learning-based method proposed to solve the SISR problem was SRCNN (27). The authors proposed a deep convolutional neural network (CNN) composed of three convolutional layers. The first layer extracts features from the LR image; the second is used to learn a non-linear mapping between features and HR images, and the third is used to reconstruct the HR image from features. The authors trained their network using a large dataset of images and obtained state-of-the-art results. SRCNN paved the way for a myriad of architectures that would be investigated to improve the SR task performance (28–30).

2.1.1 Evaluation metrics

In this section, the definition of the traditional metrics that have been widely used to compare the performance of SR task performance(31) are presented: the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM).

2.1.1.1 PSNR

PSNR measures the signal intensity ratio to the noise intensity, expressed in the logarithmetic Decibel scale. It is a prevalent metric to evaluate image quality as it is simple to calculate, have precise physical meanings, and is mathematically convenient in the context of optimization (32, 33).

The following formulation defines it:

$$\text{PSNR} = 20\log_{10}\frac{x}{\sqrt{\text{MSE}}} \tag{2.2}$$

where x is the maximum pixel value and MSE is the mean squared error:

$$\text{MSE} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{y}_{i,j} - y_{i,j})^2 \tag{2.3}$$

Where m and n are the height and width of the image, respectively. $\hat{y}_{i,j}$ is the value of the i-j pixel of the enhanced image, and $f_{i,j}$ is the value of the original image.

PSNR values approach ∞ as MSE gets closer to zero, showing that higher PSNR values are associated with higher image quality. However, PSNR has some limitations as an evaluation metric. First, it is not a perceptual metric, meaning it does not correlate well with the subjective quality of an image. Second, it is not robust to typical image processing operations, such as sharpening and noise reduction. Finally, PSNR is sensitive to the bit depth of the images being compared, so it is not always directly comparable between different image formats (33).

2.1.1.2 SSIM

The SSIM metric is an improvement over PSNR, as it is more robust to typical image processing operations and a more reliable measure of the perception of the human visual system (34). However, it is also not a perfect metric, as it is not a complete model of the human visual system, and it can be biased by the bit depth of the images being compared (33).

SSIM combines three relatively independent terms: luminance, contrast, and structure. The luminance term is a measure of the mean intensity of the image, and the contrast term is a measure of the contrast in the image. The structure term measures how similar the two images are and is defined as the product of the luminance and contrast terms. The structure term is defined as:

$$S(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2.4}$$

Where μ_x and μ_y are the means of the two images, and C_1 is a constant.

The contrast term is defined as:

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2.5}$$

Where σ_x and σ_y are the standard deviations of the two images, and C_2 is a constant. The luminance term is defined as:

$$L(x,y) = \frac{\mu_{x'}\mu_{y'} + C_3}{\mu_{x'}^2 + \mu_{y'}^2 + C_3} \tag{2.6}$$

Where $\mu_{x'}$ and $\mu_{y'}$ are the means of the two images after they have been normalized by their standard deviations, and C_3 is a constant. The constants C_1, C_2, and C_3 are used to stabilize the division with weak denominator values.

The SSIM index is then defined as:

$$SSIM(x,y) = (L(x,y))^{\alpha} \cdot (C(x,y))^{\beta} \cdot (S(x,y))^{\gamma} = L(x,y)^{\alpha} C(x,y)^{\beta} S(x,y)^{\gamma} \tag{2.7}$$

The SSIM index is a value between 0 and 1, where values close to 1 represent similar images and values close to -1 represent very dissimilar images.

2.2 Deep learning

Deep learning is a branch of machine learning techniques used to learn high-level abstractions from data using Artificial Neural Networks (ANNs). Its history can be traced back to 1943 when Pitts and McCulloch created an algorithm inspired by the human brain's neural structure. However, it was only after 2006, when Geoffrey Hinton introduced the concept, that it became a prominent research topic (35). Since then, it has dragged much attention from researchers and practitioners alike.

Part of its popularity is because it is a powerful class of computational algorithms. The Universal Approximation Theorem (36) implies that ANNs can approximate any continuous function by applying a single hidden layer of neurons with increased precision as more neurons are added.

Artificial Neural Networks can be defined as a set of interconnected processing nodes, where each node performs a simple mathematical operation on its input. The output of each node is then passed to the next node, and the final output is the result of the computation performed by the entire network.

The nodes in an Artificial Neural Network are typically arranged in layers of neurons. The first layer is the input layer, which receives the input data. The last layer is the output layer, which produces the network's output. The layers in between are called hidden layers.

Neurons are the basic processing units in an Artificial Neural Network. They are connected in a directed graph, where the edges represent the connections between the neurons. The simplest example of a neuron is called a Perceptron and is generally defined as:

$$f(x) = h(W \cdot x + e) \tag{2.8}$$

Where x is the input data, g is the activation function, W is the weight vector, and e is the error. The Perceptron algorithm is limited because its solution space is restricted to those that present linear separability. Thus, more complex deep learning algorithms have been designed to extend the solution space, such as Multi-Layer Perceptrons

(MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs), to name a few.

All deep learning algorithms have in common that they learn the weights of the connections between the neurons in the network using a training data set. The training set is a set of examples used to refine the network, adjusting the weights so the network can learn to map the inputs to the correct outputs.

For the current work, it is crucial to highlight deep learning's contribution to the achievement of state-of-the-art performance in different tasks in the medical imaging field, like segmentation and registration, automatic labeling, and lesion detection and diagnosis. For instance, Shin (*37*) achieved state-of-the-art performance on the mediastinal lymph node detection by applying deep convolutional neural networks; Han (*38*) demonstrated that Alzheimer's Disease could be reliably detected at a very early stage with the application of an unsupervised method based on a Deep Generative Adversarial Network. You . (*12*) proposed a GAN-based image-super resolution method that quantitative and qualitatively outperformed conventional state-of-the-art image enhancement techniques; Miao (*39*) proposed an approach based on Convolutional Neural Networks to register 2D and 3D medical images in real-time 2D/3D with a demonstrated high accuracy.

In the next section, *Generative Neural Networks* will be described in deeper detail, along with the main architectures applied for single image super-resolution.

2.2.1 Generative Adversarial Networks

Imagine a currency counterfeiter trying to forge fake currency and using it without being detected, whereas the police try to discriminate which currency is real or counterfeit. The constant competition makes the parties level their game until the counterfeit currency is indistinguishable from the real ones. Goodfellow (*40*) compared the framework of *Generative Adversarial Networks* (GAN) to this competition.

The vanilla GAN implementation consists of two models trained simultaneously: G, the generator (the counterfeiter in the example), and D, the discriminator (the police). The role of G is to construct samples that are the most similar to the training data distribution; D, however, must estimate the probability of a sample coming from the training sample rather than the model G. We can generically express the objective function of the minimax game played by G and D as:

$$\min_G \max_D \mathcal{V}(G,D) = \mathcal{L}(G) + \mathcal{L}(D) \tag{2.9}$$

A diagram of this model is shown in Fig. 2.1. In this vanilla implementation, the generator is a differentiable function represented by a Multi-Layer Perceptron with parameters and noise as input; the discriminator is a second Multi-Layer Perceptron with

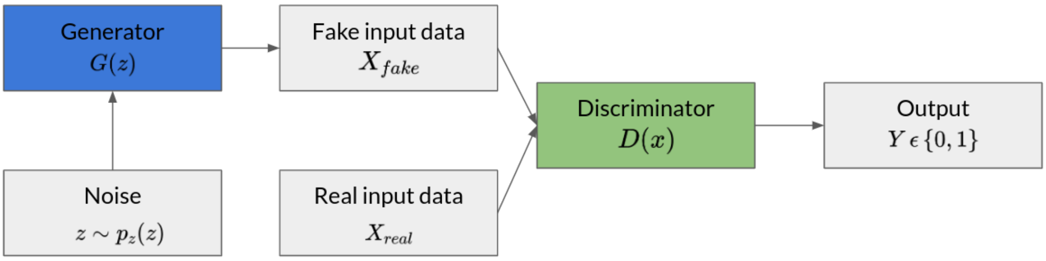

Figure 2.1: **Vanilla GAN diagram**

parameters and the data sample as input. In this implementation, the optimization function is the Log-likelihood. Thus we can replace $\mathscr{L}(G)$ and $\mathscr{L}(D)$:

$$\min_{G}\max_{D}\mathscr{V}(G,D) = \mathbb{E}_{x\sim\mathbb{P}_x}[\log 1 - D(G(z))] + \mathbb{E}_{x\sim\mathbb{P}_x}[\log D(x)] \tag{2.10}$$

Since its debut, GANs have been applied to a wide range of domains, with promising results in many different applications, like the text to image synthesis, data compression, super-resolution of images, image enhancement, style transfer, image-to-image translation, video generation, and more. However, despite the increasing attention it has received in the last years, a few challenges in training GAN models are still remarkable.

In the following sub-sections, relevant GAN models to substantiate the theoretical framework in the methodology presented in Chapter 3 are expanded upon.

2.2.2 WGAN and WGAN-GP

As G and D play a minimax game, its optimal solution is found when both the generator and the discriminator cannot unilaterally improve their losses. This state is called a Nash equilibrium. However, gradient descent may fail to converge because it is a local optimization method, leading only to a local Nash equilibrium (*41*). GANs have also been criticized for their inability to learn the whole data distribution, which leads to a scenario called mode collapse, or Helvetica. Mode collapse happens when the generator gets stuck mapping different inputs to the same output, resulting in a lack of diversity and poor generalization (*42*). Besides, there is also no consensus as to which measure should be used for a fair model comparison (*43*). Several GAN variants have been proposed to deal with these problems. For instance, in WGAN paper (*44*) proposed solving the mode collapse problem by using the Wasserstein-1 distance to optimize the network and clipping the discriminator's weights to enforce a Lipschitz constraint on the discriminator gradient. Thus, the discriminator is forced to have a maximum gradient, which results in a more stable optimization process and an overall loss correlated with the convergence of the generator.

The WGAN objective function can be expressed as:

$$\min_{G}\max_{D\in\mathscr{D}}\mathscr{V}(G,D)=\mathbb{E}[D(x)]-\mathbb{E}[D(G(z))] \tag{2.11}$$

Where $\mathscr{D}$ is the set of functions that satisfy the Lipschitz constraint. Weight clipping, however, may bias the discriminator towards simple functions, causing it to become too weak, leading to undesired behaviors of the networks, like overfitting the generator. In the WGAN-GP paper, to solve this problem, (45) proposed penalizing the norm of the discriminator's gradient with respect to its input, which forces the generator to spread out its support, avoiding mode collapse.

The WGAN-GP objective function can be expressed as:

$$\min_{G}\max_{D\in\mathscr{D}}\mathscr{V}(G,D)=\mathbb{E}[D(x)]-\mathbb{E}[D(G(z))]+\lambda\mathbb{E}[(\|\nabla_{\hat{x}}D(\hat{x})\|_2-1)^2] \tag{2.12}$$

where λ is the penalty coefficient, and $\|\nabla_{\hat{x}}D(\hat{x})\|_2$ is the l_2 norm of the gradient of $\hat{x}$, the interpolation between the real input x and the generated input $G(z)$.

2.2.3 Pix2pixGAN

Isola (46) demonstrated that Generative Adversarial Networks are also a promising approach to solving image-to-image translation problems. Pix2pix GAN, presented in the paper, is a GAN trained with paired images, which conditions the generator output on an input image. This conditional GAN learns a mapping from image X to image Y by training a generator G to translate X to Y conditioned on X, simultaneously training an adversary D to distinguish between translated images $G(X)$ and actual images Y. Its objective function in this method is given by:

$$\mathscr{L}_{cGAN(G,D)}=\mathbb{E}[\log D(x,y)]+\mathbb{E}[\log(1-D(G(x,z)))] \tag{2.13}$$

These changes in the loss function encourage the generator to learn a mapping from image x and noise vector z to y, i.e., $G:x,z\rightarrow y$. The noise vector z is inserted because the generator could fail to converge and get stuck in a deterministic mapping function without it. Besides, the authors also proposed the addition of an L1 distance to the total loss in order to task the generator to be closer to the ground truth in the L1 sense, encouraging solutions with less blur:

$$\mathscr{L}_{L1}=\mathbb{E}[\|y-G(x,z)\|_1] \tag{2.14}$$

Thus the final objective function of pix2pix GAN is given by:

$$L_{pix2pix}=\mathscr{L}_{cGAN(G,D)}+\lambda\mathscr{L}_{L1} \tag{2.15}$$

Where λ is the penalty coefficient, regulating the importance of $\lambda\mathscr{L}_{L1}$. The game between the generator and the discriminator is played until the discriminator can no

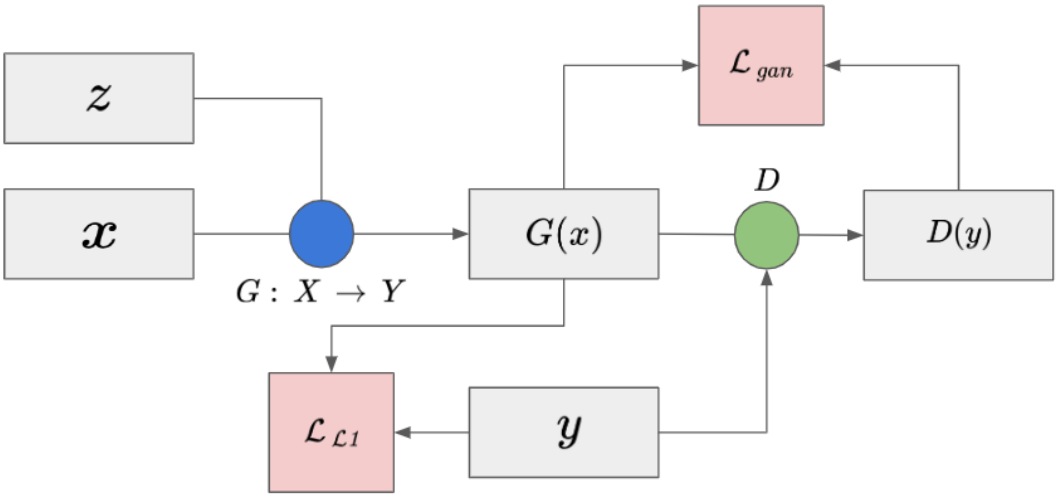

Figure 2.2: **Pix2pix GAN diagram**

longer distinguish between actual and translated images. A diagram of this model is shown in Fig. 2.2.

2.2.4 CycleGAN

Only the adversarial loss, as seen in pix2pix GAN, cannot ensure that learned functions will map a given input to a desired output in the target domain. Thus, Zhu et al. proposed (*10*) adding a Cycle Consistency Loss to the objective function to enable the model to translate images from different domains without paired training examples, which can also prevent degeneracy in the learning process.

The cycle consistency property assumes that if we have a translator F that maps elements from domain X to Y, i.e., $F : X \rightarrow Y$, and we have another translator G that maps elements from domain Y to X, i.e., $G : Y \rightarrow X$, then F and G should be inverses of one another, i.e., $F(G(x)) \approx x$ (forward cycle consistency) and $G(F(y)) \approx y$ (backward cycle consistency). Thus, the following loss function is defined to encourage this behavior:

$$\mathscr{L}_{Cyc}(F,G) = \mathbb{E}_{x\sim p_x}[||F(G(x)) - x||_1] + \mathbb{E}_{y\sim p_y}[||G(F(y)) - y||_1] \tag{2.16}$$

The $l_1 - norm$ is chosen because, in experiments, it demonstrated to foster less blur in images generated. This cycle consistency loss is combined with the adversarial loss, applied to each of the generators:

$$\mathscr{L}_{GAN}(G,D,X,Y) = \mathbb{E}_{y\sim p_y}[\log D(y)] + \mathbb{E}_{x\sim p_x}[\log(1 - D(G(x))] \tag{2.17}$$

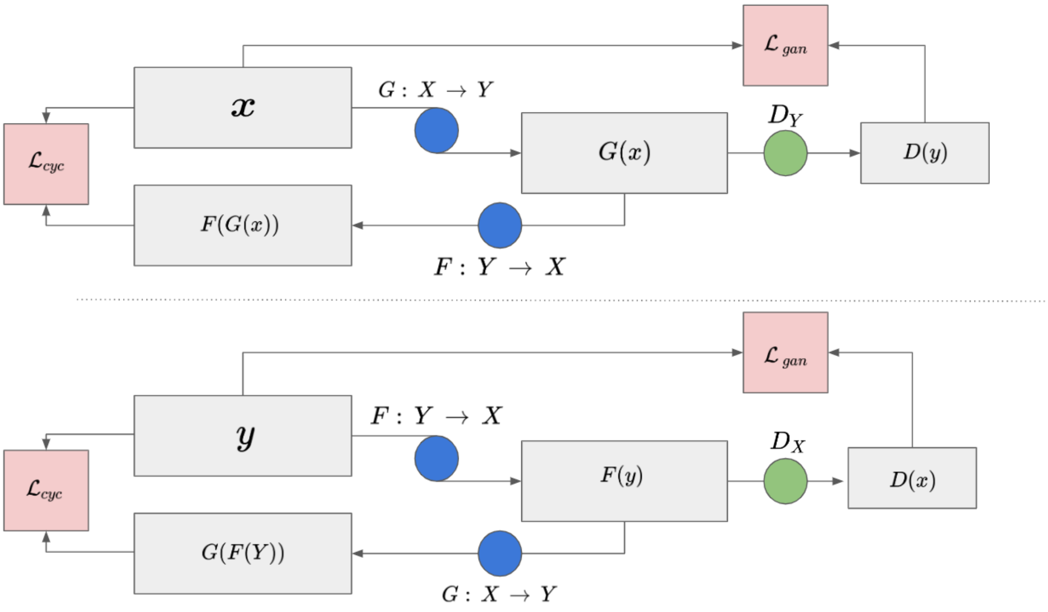

Figure 2.3: **CycleGAN diagram**

Thus we have the final objective function for the CycleGAN:

$$\mathcal{L}_{Cycle}(G,F,D_X,D_Y) = \mathcal{L}_{GAN}(G,D_Y,X,Y) + \mathcal{L}_{GAN}(F,D_X,Y,X) + \mathcal{L}_{Cyc}(F,G) \tag{2.18}$$

A diagram of this model is shown in Fig. 2.3.

2.2.5 CinCGAN

Expanding the framework of CycleGAN (*10*), Yuan (*47*) proposed a Cycle-in-Cycle GAN, named CinCGAN, to approach the super-resolution task as an image-to-image translation problem.

One of the limitations in super-resolution tasks with Deep Learning algorithms is the dependence on low and high-resolution paired images. In CycleGAN, images are translated from different domains with unpaired training data. However, one of the limitations in using this model to solve a super-resolution problem is the assumption that the input and output image has the same size. In super-resolution problems, however, the output can be much larger than the input.

In CinCGAN, (*47*) approaches this problem by combining two CycleGANs: the first maps the low-resolution image to a bicubic-downsampled low-resolution image. This module has denoising and deblurring role in the architecture. The output is then up-sampled and fed into the second one, encapsulating the first one. Thus, the solution is presented in three steps: first, the low-resolution image is mapped to a bicubic-downsampled image ($X \to Y$); second, the bicubic-downsampled image is mapped to a high-resolution one through an existent super-resolution model ($Y \to Z$); and third, both models are combined to get the final output. Thus, mapping images from the low-resolution (LR) domain to the high-resolution (HR) domain can be done with unpaired training data and without the same shape assumptions.

The key idea is to allow the two mappers to share information. The first mapper learns the low-level image structure, while the second mapper learns the high-level image structure. By sharing information, the two mappers can learn a complete image representation, which is then used to generate the final high-resolution image.

The framework diagram of this model is shown in 2.4

In this model, negative log-likelihood loss calculated in the Adversarial Loss of CycleGAN is replaced by a least square loss. Thus, Adversarial Loss can be defined as follows for the first network and the second, respectively:

$$\mathscr{L}^1_{GAN} = \mathbb{E}_{x \sim p_x}[||D_1(G(x)) - 1||_2] \tag{2.19}$$

$$\mathscr{L}^2_{GAN} = \mathbb{E}_{x \sim p_x}[||D_2(SR(G(x))) - 1||_2] \tag{2.20}$$

The Cycle Consistent Loss is similar to the formulation in (*10*), except that (*47*) chose the $l_2 - norm$ over the $l_1 - norm$, defined for the first network and the second, respectively:

$$\mathscr{L}^1_{Cyc} = \mathbb{E}_{x \sim p_x}[||F(G(x)) - x||_2] \tag{2.21}$$

$$\mathscr{L}^2_{Cyc} = \mathbb{E}_{x \sim p_x}[||H(SR(G(x))) - x||_2] \tag{2.22}$$

In order to avoid color variation in the first network, CinCGAN also adds an Identity Loss to the objective function, which is defined as:

$$\mathscr{L}^1_{Idt} = \mathbb{E}_{x \sim p_x}[||G(y)) - y||_1] \tag{2.23}$$

For the second network it is slightly different, as instead of encouraging color consistency, Identity Loss encourages super-resolution consistency:

$$\mathscr{L}^2_{Idt} = \mathbb{E}_{x \sim p_x}[||SR(z\prime)) - z||_1] \tag{2.24}$$

Where $z\prime$ is the high-resolution image downsampled with a bicubic kernel. In order to avoid artifacts insertion in the output image, CinCGAN also adds a Total Variation Loss, encouraging spatial smoothness in the generated images:

$$\mathscr{L}_{\mathrm{TV}}(G) = \sum_i \sum_{j \in \{1,2\}} \left(\left| \frac{\partial G}{\partial x_{i,j}} \right| + \left| \frac{\partial G}{\partial y_{i,j}} \right| \right), \tag{2.25}$$

In which $x_{i,j}$ and $y_{i,j}$ are the spatial coordinates of the generated image, and G is replaced by $SR(G)$ for the second network.

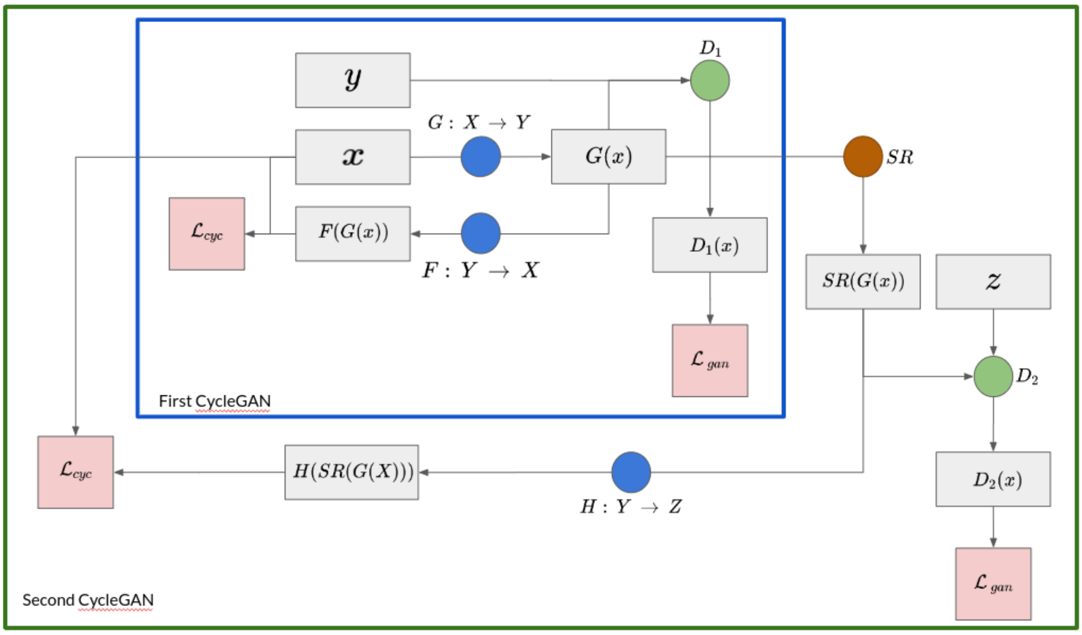

Figure 2.4: **Cycle-in-Cycle GAN diagram**

Thus we have the final objective function, defined for the first network and the second, respectively:

$$\mathscr{L}_{CinC-1} = \mathscr{L}^1_{\text{GAN}} + \lambda_1 \mathscr{L}^1_{\text{Cyc}} + \lambda_2 \mathscr{L}^1_{\text{Idt}} + \lambda_3 \mathscr{L}^1_{\text{TV}} \tag{2.26}$$

$$\mathscr{L}_{CinC-2} = \mathscr{L}^2_{\text{GAN}} + \lambda_1 \mathscr{L}^2_{\text{Cyc}} + \lambda_2 \mathscr{L}^2_{\text{Idt}} + \lambda_3 \mathscr{L}^2_{\text{TV}} \tag{2.27}$$

Where λ is a weighting parameter chosen to balance the different losses. Then, the first CycleGAN network is pre-trained to learn the $X \to Y$ mapping. After it converges, the second CycleGAN network is fine-tuned, training jointly to learn the mapping $X \to Z$.

2.2.6 GAN-CIRCLE

Generative Adversarial Networks (GANs) have also been increasingly exploited for medical image super-resolution (12, 13, 18, 48–52).

The most prominent work on CT image SR is GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE)(12), outperforming previous works(13–16). GAN-CIRCLE can preserve anatomical information and suppress noise, leading to an excellent diagnostic performance in terms of traditional image quality metrics(12, 17). For example, Guha (17) exploited GAN-CIRCLE to super-resolve trabecular bone micro-structures and improved the structural similarity index (SSIM).

Expanding the CycleGAN framework, GAN-CIRCLE proposes a method to enforce the cycle consistency in terms of the Wasserstein distance. Thus, the negative log-likelihood in the Adversarial Loss is replaced by the Wasserstein distance and is defined as follows, in both directions:

$$\mathscr{L}_{\mathrm{GAN}(D_Y,G} = -\mathbb{E}_y[D(y)] + \mathbb{E}_x[D(G(x)] + \lambda \mathbb{E}_{\hat{y}}[(\|\nabla_{\hat{y}} D(\hat{y})\|_2 - 1)^2] \tag{2.28}$$

$$\mathscr{L}_{\mathrm{GAN}(D_X,F} = -\mathbb{E}_x[D(x)] + \mathbb{E}_y[D(F(y)] + \lambda \mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{2.29}$$

Where, similar to Eq. 2.12, the first two terms define the loss in terms of the Wasserstein distance, and the third one is the $l_2 norm$ of the gradient, added to enforce the Lipschitz continuity property, and $\hat{y}$ is the interpolation between the actual HR image y and the generated output $G(x)$.

GAN-CIRCLE also adds an Identity Loss term to the model objective function, similar to the Eq. 2.23:

$$\mathscr{L}_{Idt}(G,F) = \mathbb{E}_y[\|G(x)) - x\|_1] + \mathbb{E}_x[\|G(x)) - x\|_1] \tag{2.30}$$

In order to promote image sparsity and reduced noise, GAN-CIRCLE also implements a Joint Sparsifying Transform Loss term, based on the Total Variation Loss, adding a second component in a non-linear combination in order to encourage the minimization of the difference image $y - G(x)$ and thus preserving anatomical characteristics:

$$\mathscr{L}_{\mathrm{JST}}(G) = \tau \mathscr{L}_{\mathrm{TV}}(G(x)) + (1 - \tau)\mathscr{L}_{\mathrm{TV}}(y - G(x)) \tag{2.31}$$

where τ is a scaling factor and $\mathscr{L}_{\mathrm{TV}}$ is the same as Eq. 2.25.

Thus, the GAN-CIRCLE loss function combines four different loss terms to regularize the training procedure by enforcing the desired mappings:

- an *adversarial loss term* ($\mathscr{L}_{\mathrm{Adv}}$) to enforce the matching of empirical distributions in the source and target domains;

- a *cycle-consistency loss term* ($\mathscr{L}_{\mathrm{Cyc}}$) to prevent degeneracy in the adversarial learning and promote forward and backward cycle consistency, defined as $G(F(\mathbf{I}_{\mathrm{hr}}) \approx \mathbf{I}_{\mathrm{hr}}$ and $F(G(\mathbf{I}_{\mathrm{lr}})) \approx \mathbf{I}_{\mathrm{lr}}$;

- an *identity loss term* ($\mathscr{L}_{\mathrm{IDT}}$) to regularize the training process and promote the relationships $G(\mathbf{I}_{\mathrm{hr}}) \approx \mathbf{I}_{\mathrm{hr}}$ and $F(\mathbf{I}_{\mathrm{lr}}) \approx \mathbf{I}_{\mathrm{lr}}$;

- a *joint sparsifying loss term* ($\mathscr{L}_{\mathrm{JST}}$) to promote image sparsity and reduced noise.

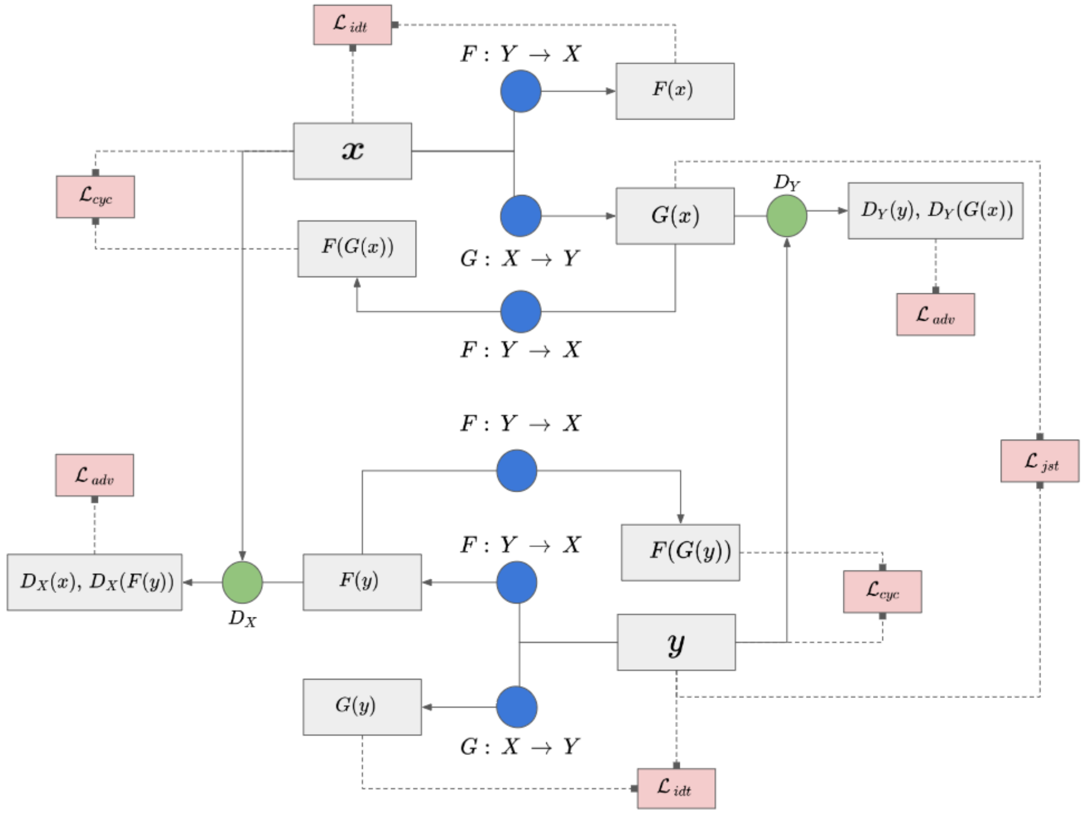

Figure 2.5: **GAN-CIRCLE framework diagram**

The overall loss function used for training is defined as:

$$\mathscr{L}_{\text{CIRCLE}} = \mathscr{L}_{\text{Adv}}(D_{\text{HR}}, G) + \mathscr{L}_{\text{Adv}}(D_{\text{LR}}, F) + \lambda_1 \mathscr{L}_{Cyc}(G, F) + \lambda_2 \mathscr{L}_{\text{IDT}}(G, F) + \lambda_3 \mathscr{L}_{\text{JST}}(G), \tag{2.32}$$

where λ_1, λ_2 and λ_3 are weighting parameters to balance the different loss terms, respectively. The framework diagram of this model is shown in 2.4

2.3 Radiomics

Gillies (1) described radiomics as "the process of extracting a large number of features from medical images, to provide quantitative descriptions of the image content, to be used as inputs to computational models or as biomarkers". These features are subsequently stored, and the data is mined to generate research hypotheses and develop tools to support medical decisions.

Typically, radiomics feature extraction involves three main steps: image preprocessing, segmentation, and feature extraction.

- The preprocessing step is optional and may be used to improve the segmentation performance or standardize features.
- In the segmentation step, the Region of Interest (ROI) extraction is performed. The voxels of a tumor, for instance.
- Then, the features are extracted from the ROI, thus translating the image into semantic and agnostic features of the region of interest. Thus enabling the quantitative measurement of intra and intertumoral heterogeneity, for example.

According to Lambin (53), the radiomic hypothesis is that genomic and proteomic patterns can be expressed in terms of image characteristics. This hypothesis is supported by image-guided biopsies, which demonstrated that protein expression patterns within a tumor are associated with spatial differences.

For example, a study conducted in 2017 by Vargas (54) identified the clinical and biological validity of several preoperative radiomic features significantly associated with time-to-disease progression (TTP) and classification of ovarian cancer (CLOVAR) gene expression profile. Another interesting finding reported in the literature demonstrates that more tumors with more genomic heterogeneity are more likely to develop a resistance to treatment and metastasize. Considering the Radiomics hypothesis described above, this heterogeneity could be assessed through the spatial differences in medical imaging. Segal (55) showed that the combination of 28 imaging traits would suffice to reconstruct the variation of 116 gene expressions in hepatocellular carcinomas.

Radiomics features can be classified into semantic and agnostic features.

Semantic features are features that have a known, specific meaning in the field of medicine and usually describe the ROI — for example, the average tumor diameter, the maximum tumor diameter, or the tumor volume. The main advantage of these features is that radiologists can easily understand them. **Agnostic features**, on the other hand, are not guaranteed to be clinically relevant. The main advantages of agnostic features are that they are more robust because they do not depend on the tumor's specific characteristics and may be more efficient in predicting clinical outcomes. They describe the region of interest heterogeneity through quantitative descriptors commonly divided into first, second, and higher-order statistical outputs. Some examples are the entropy, the inverse difference moment, and the variance.

First-order statistics describe the distribution of values of individual voxels without caring for spatial relationships, like histogram-based methods (mean, median, max, min).

Second-order statistics, or texture features, describe statistical relationships between voxels according to their similarity of values. These measurements were first introduced by Haralick in 1973 (56). Second-order statistics, also referred to as texture features, provide both intensity and spatial information. They describe the distribution of voxel intensity values between neighboring voxels along with different directions and distances and are derived from so-called gray-tone-spatial-dependence matrices.

Finally, higher-order statistics apply filter grids on the image to extract repetitive or unique patterns. These features include fractal analyses, Minkowski functionals, wavelets, Laplacian transformations, and Gaussian bandpass filters.

Despite its great potential to enhance precision medicine, Radiomics has not been fully integrated into clinical practice for several reasons. In the first place, radiomics is a complex process that requires a high level of expertise in image processing and analysis. In addition, radiomics features do not directly translate into clinical outcomes, but need to be integrated into a model to be interpretable. There is also the question of standardization, the process of feature extraction and evaluation, and comparing results from different studies. Finally, the lack of hardware and software integrations hampers the clinical adoption of radiomics.

2.3.1 Feature robustness analysis

Though there is no straightforward definition of radiomic feature robustness, Jha *et al.* (57) proposed a feature to be classified as robust when it is stable (has low variability) under changing conditions. This is stability is define in terms of two concepts: repeatability and reproducibility of radiomic features. Repeatability refers to features that remain the same when imaged multiple times in the same subject, using the same image acquisition methods. Reproducibility refers to features that remain the same when extracted using different equipment, different software, different image acquisition settings, or different operators.

One of the most challenging problems for radiomic models to be implemented in a clinical setting is related to the robustness of these models when different datasets are used. Patient position in the image acquisition process, imaging parameters, and ROI segmentation have different impacts on the radiomic features. Using features that are not robustly stable against the perturbations will cause a radiomic model to perform poorly when used to generate predictions on unseen data, making them impracticable to be applied in a clinical setting. Thus, assessing feature robustness is essential to improve model generalisability and clinical application.

Shafiq Ul Hassam and coauthors(6) studied the impact of slice thickness and pixel spacing on radiomic features extracted from computed tomography phantom images acquired with different scanners as well as different acquisition and reconstruction parameters. Experimental results demonstrated that voxel size resampling is an appropriate preprocessing step for image data sets acquired with variable voxel sizes, and it allows for obtaining more reproducible features. Additionally, while some radiomic features were voxel size and gray-level discretization dependent, the use of normalizing factors in their definitions may reduce or remove such dependencies. These findings were validated in a subsequent work(58), where the voxel size and gray levels in phantom normalizations are applied in lung tumors images. Based on the results obtained by considering eighteen patients with non-small cell lung cancer of varying tumor volumes, the authors concluded that voxel size and gray-levels normalizations improve the robustness of radiomic features for lung tumor images.

More recently, Sanchez (7) investigated the robustness of radiomic features in computed tomography images with different slice thicknesses for liver tumors and muscle. After addressing the dependencies of texture radiomic features by choosing the optimal number of gray levels, features were compared across thicknesses to identify reproducible features. The study considered a computed tomography dataset of 43 patients with hepatocellular carcinoma, and the analysis showed consistent results for both tumor and muscle tissue. In particular, high robustness of a large fraction of features (75 – 90%) was found, thus allowing the authors to define guidelines for radiomic studies using variable slice thickness.

Le (8) investigated how image resampling (involving interpolation) and perturbations on the regions of interest (ROIs) affect the robustness of the features. They extracted 93 radiomic features from carotid artery CT angiograms of 41 patients with cerebrovascular events. Radiomic feature robustness was assessed against region-of-interest perturbations, image preprocessing settings, and quantization methods using single- and multi-slice approaches. The analysis showed that, by proving in input to machine learning algorithms the most robust features, it is possible to identify the culprit and non-culprit arteries. Multi-slice features were superior to single for producing robust radiomic features, and the optimal image quantization method used bin widths of 25 or 30. The results suggest introducing carotid computed tomography radiomics into clinical practice to improve stroke prediction and target therapies for those at the highest risk. In the same vein, Mottola (5) investigated feature reproducibility against noise, varying resolutions, and segmentations in a computed tomography dataset of 98 renal cell carcinomas and 93 contralateral normal kidneys. Experimental results highlighted the importance of the interpolation method, with the Lanczos interpolation being the most effective at preserving original information in resampling.

Zwanenburg (4) considered 18 methods to determine feature robustness based on image perturbations. Experimental results, considering 4032 features, showed that a perturbation chain consisting of noise addition, affine translation, volume growth/shrinkage, and contour randomization identified the fewest false-positive robust features. Thus, this perturbation chain may represent a viable option to evaluate feature robustness.

In order to measure the feature robustness, two statistical measures are commonly used: the spearman rank coefficient and the Intraclass Correlation.

2.3.1.1 Spearman Rank Coefficient

The Spearman rank correlation coefficient measures the strength and direction of the relationship between two variables. The value of ρ can range from -1 to 1, where -1 indicates a perfect negative monotonic relationship and 1 indicates a perfect positive monotonic relationship. The value of ρ is 0 if there is no monotonic relationship between the two variables.

The Spearman rank correlation coefficient is a non-parametric measure of association, as it does not assume that the variables are normally distributed. The Spearman rank correlation coefficient is also sensitive to outliers, so it is essential to be aware of them when interpreting the results.

The formula for the Spearman rank correlation coefficient is:

$$\rho = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(s_x)(s_y)} \tag{2.33}$$

Where ρ is the Spearman rank correlation coefficient, n is the number of pairs of data, x_i is the rank of the $i-th$ data point in the x variable, $\bar{x}$ is the mean rank of the x variable, y_i is the rank of the ith data point in the y variable, $\bar{y}$ is the mean rank of the y variable, s_x is the standard deviation of the ranks of the x variable, and s_y is the standard deviation of the ranks of the y variable.

Traverso (59) reports in a systematic review of the usage of Spearman Rank Coefficient as one of the primary metrics used to measure feature robustness.

2.3.1.2 Intraclass Correlation (ICC)

The ICC is a statistical measure describing the correlation and agreement between measurements and is especially adequate when a high correlation is expected within a specific class. This metric is the most commonly reported in robustness studies (59).

The ICC calculations are performed separately for each feature and perturbation studied. Zwanenburg (60) classifies features with ICC 0.90 as robust, which keeps the error below 0.05 even for a small patient cohort (61).

The ICC is generically calculated as:

$$ICC = \frac{SS_{between} - SS_{within}}{SS_{between} + SS_{within}}, \tag{2.34}$$

where $SS_{between}$ is the sum of squares between groups, SS_{within} is the sum of squares within groups, and SS_{total} is the total sum of squares. The interpretation of the ICC metric depends on the type of study. There are ten versions of the ICC with different assumptions and interpretations, four of which are the most commonly used in robustness studies: ICC(1,1) for single-measurement agreements, ICC(2,1) for two-measurement agreements, ICC(3,1) for three-measurement agreement, and ICC(3,k) for k-measurement agreement. The ICC(1,1) is the simplest and most suitable for the first two cases. The ICC(3,1) is the most suitable for the last two cases (62).

Chapter

3 Materials and Methods

3.1 Datasets and Preprocessing

DeepLesion dataset

The DeepLesion dataset(63) contains 32,120 CT images of different types of lesions in different parts of the body from 4,427 unique patients, along with accompanying 2D diameter measurements and bounding-boxes of lesions and semantic labels.

This dataset was used specifically for training and evaluating the GAN-CIRCLE network used for the Super-Resolution task. For training, 10,000 randomly selected CT images with an image size of 512 × 512 pixels and in-plane pixel spacing between 0.18 and 0.98 mm (median: 0.82 mm) were selected from the available images. Out of these, 1,000 CT images were randomly held-out in order to assess further model performance.

NSCLC-Radiomics dataset

The Non-Small Cell Lung Cancer-Radiomics (NSCLC-Radiomics) dataset(64) is a well-established publicly available dataset that contains CT slices from 422 NSCLC patients. This dataset is available *via* The Cancer Imaging Archive (TCIA)(65). For careful and reliable radiomic analyses, our study uses a highly homogeneous subset composed of 142 CT scans, accounting for 17,938 CT slices with an image size of 512×512 pixels, in-plane pixel spacing of 0.98 mm, and slice thickness of 3.00 mm. The B19f convolution kernel was applied on all the scans for CT image reconstruction.

The dataset provides annotated 3D tumor segmentation masks and clinical outcome data. The images are used to assess our proposed lesion-focused CIRCLE-GAN framework in terms of radiomic feature robustness.

Data preprocessing

For all the implemented SR approaches, the range of intensity for raw CT volumes was clipped to $[-100,400]$ Hounsfield Units (HU), and then normalized to $[0,1]$. We generated the Low-Resolution CT (LRCT) counterparts from the High-Resolution CT (HRCT) images by degrading them through a Gaussian white noise process with a standard deviation of 0.25 and a Gaussian blur, with a kernel size of 8×8 pixels and a bandwidth of 1.6. Afterwards, the images were downsampled with a scale of 2 and upsampled using the nearest neighbor interpolation, according to You *et al.*(12). The upsampling step improves feature extraction by enforcing the same image size for LRCT and HRCT(66).

Image patches were then cropped based on the lesion bounding box annotations in the metadata—the cropping process leads to avoiding artifact generation out of the lesion area(18). The preprocessing pipeline is displayed in Fig. 3.1. The bounding box cropping was performed on the image before the degradation process to obtain the patch that was treated as the ground-truth HRCT image and also after the degradation process, to obtain its LRCT patch counterpart. This step was performed to constrain the network to focus on the regions of interest (ROI), thus reducing the synthetization of artifacts from regions beyond the lesion area(18)

By applying this procedure only on the Deeplesion dataset, we generated 10,000 LRCT/HRCT patches with similar image sizes for training a CIRCLE-GAN-based SR model. The TCIA NSCLC CT dataset was used solely for radiomic feature robustness assessment.

3.2 Lesion focused GAN-CIRCLE-based image Super-Resolution

3.2.1 Network architecture

A modified version of GAN-CIRCLE(12) was implemented to tackle the SR problem. The GAN-CIRCLE is a cycle-consistent adversarial model consisting of two non-linear generative mappings and their respective discriminators that are trained jointly for optimal convergence, as described in Ch.2.

The generator networks G and F share the same architecture, which is composed of a feature extraction and a reconstruction network. The *feature extraction network* consists of twelve layers (i.e., feature blocks) of 3×3 convolution kernels, bias, Leaky Rectified Linear Unit (ReLU) activation, and dropout. Each block output is concatenated through skip connections before the reconstruction network to capture local/global image features. The number of output filters in each convolutional layer is set according to You (12). In the *reconstruction network*, two branches are stacked in a network-in-network fashion to increase non-linearity and potentially reduce the filter space

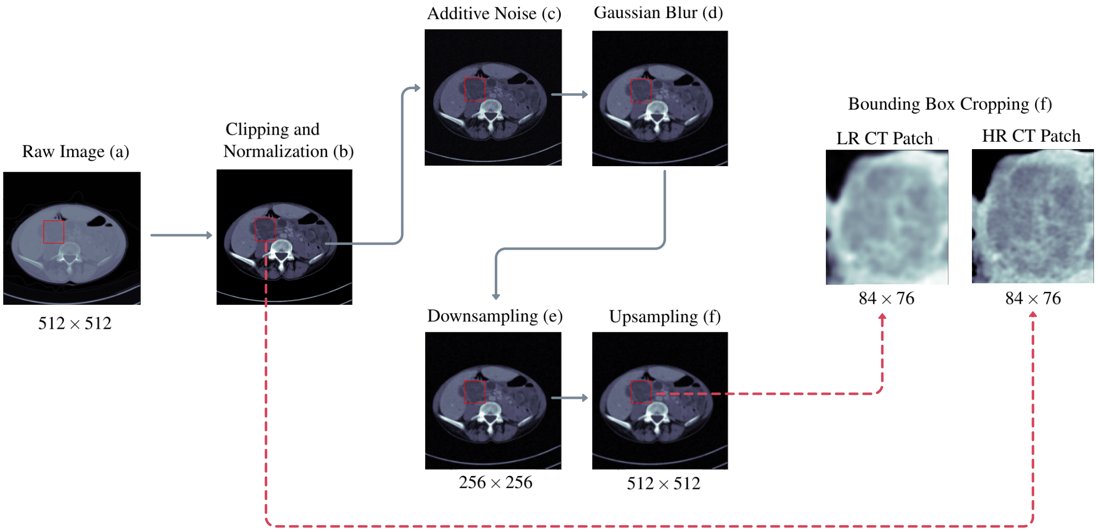

Figure 3.1: **CT image preprocessing pipeline for GAN training** The HU values of input CT images **(a)** were clipped to the range $[-100, 400]$ HU and normalized to the unit range $[0, 1]$ **(b)**. To generate the low resolution CT image counterpart, the image was perturbed by noise addition **(c)** and Gaussian blurring **(d)**, downsampled by a factor of 2× **(e)** and then upsampled to the original dimension **(f)** using a nearest neighbor interpolation method. Finally, the HRCT patch and LRCT patch were extracted from the lesion bounding box crops **(g)**.

dimension for faster computation. A transposed convolutional layer is adopted for up-sampling and the last convolutional layer combines all feature maps to produce the SR output.

One of the GAN-CIRCLE framework limitations, however, is the input scale constraint. GAN-CIRCLE (12) was trained with 32×32 low-resolution patches and 64×64 high-resolution patches, which restrains the network to be trained with ROIs, that have multi-scales. The fixed input-scale restrains the possibility to train or fine-tune the network with lesion (ROI) patches, which are varying in scale by nature. Training a network only with ROI images could be significant in cases where lesion-focused applications are in mind. For instance, Zhu *et al.* (67) proposed a lesion focused super-resolution architecture, imitating the clinicians' scrutinization procedure, i.e. focused on the ROI. The proposed network was composed by a lesion detection module, that received a MRI image and resulted in the predicted ROI, and then passed this predicted ROI as an input to the super-resolution GAN module. According to their findings, this architecture reduces significantly the cost of training the GAN for super-resolution, results in better perceptual qualities of the generated images, with less artifacts sythesised, as regions excluded from the ROI are not considered in the training process.

Inspired by Zhu *et al.* (67), and He *et al.*(68), we included a SPP layer in order to extract the features from multi-sized LRCT/HRCT input patches, allowing for the

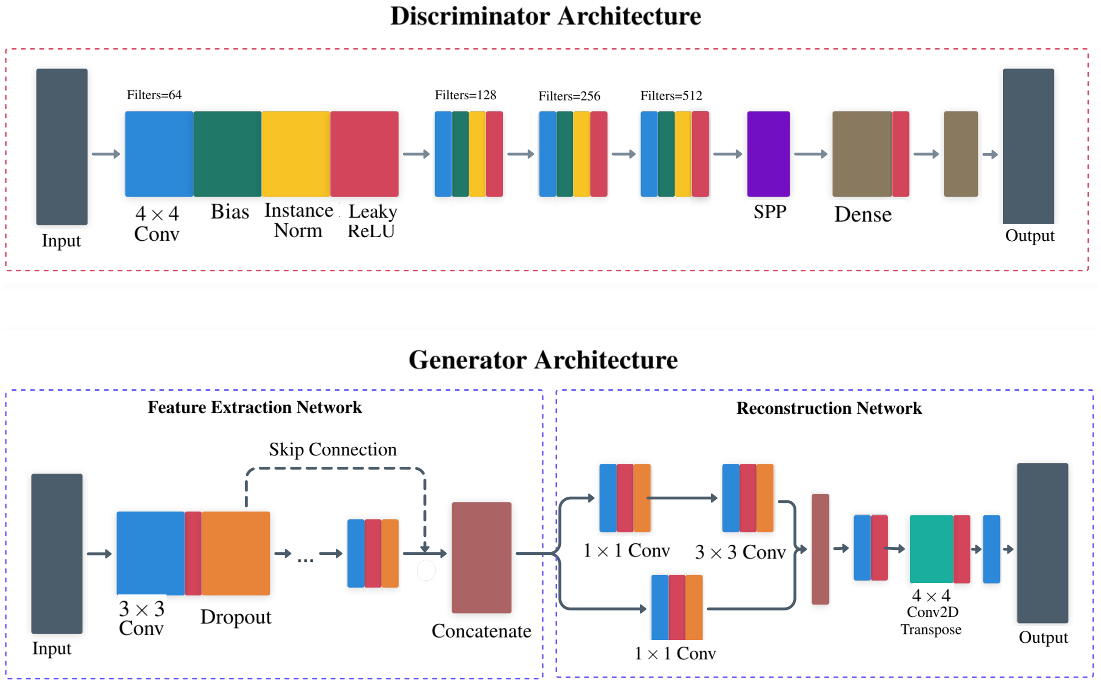

Figure 3.2: **The discriminator and generator architectures devised for GAN-SR of medical images.**

training of a lesion patch-focused network.The SPP layer consists of two parts: one is the max-pooling layer and the other is the mapping layer. The max-pooling layer is used to extract the features from the input patches, and the mapping layer is used to make the extracted features fit the layers of the discriminator.

Thus, the discriminators D_{HR} and D_{LR} also share the same network architecture, which is composed of four blocks of 4×4 convolution kernel, bias, instance normalization, and Leaky ReLU activation followed by an SPP layer and then two dense layers.

Fig. 3.2 displays the discriminator and generator architectures used in our work.

Similar to GAN-CIRCLE(*12*), the loss function also combines four different loss terms to regularize the training procedure by enforcing the desired mappings:

- an *adversarial loss term* ($\mathscr{L}_{\mathrm{Adv}}$) to enforce the matching of empirical distributions in the source and target domains;
- a *cycle-consistency loss term* ($\mathscr{L}_{\mathrm{Cyc}}$) to prevent degeneracy in the adversarial learning and promote forward and backward cycle consistency, defined as $G(F(\mathbf{I}_{\mathrm{hr}}) \approx \mathbf{I}_{\mathrm{hr}}$ and $F(G(\mathbf{I}_{\mathrm{lr}})) \approx \mathbf{I}_{\mathrm{lr}}$;
- an *identity loss term* ($\mathscr{L}_{\mathrm{IDT}}$) to regularize the training process and promote the relationships $G(\mathbf{I}_{\mathrm{hr}}) \approx \mathbf{I}_{\mathrm{hr}}$ and $F(\mathbf{I}_{\mathrm{lr}}) \approx \mathbf{I}_{\mathrm{lr}}$;
- a *joint sparsifying loss term* ($\mathscr{L}_{\mathrm{JST}}$) to promote image sparsity and reduced noise.

Thus, the overall loss function used for training is defined as:

$$\mathscr{L}_{\text{CIRCLE}} = \mathscr{L}_{\text{Adv}}(D_{\text{HR}}, G) + \mathscr{L}_{\text{Adv}}(D_{\text{LR}}, F) + \lambda_1 \mathscr{L}_{Cyc}(G, F) + \lambda_2 \mathscr{L}_{\text{IDT}}(G, F) + \lambda_3 \mathscr{L}_{\text{JST}}(G), \tag{3.1}$$

where λ_1, λ_2 and λ_3 are weighting parameters to balance the different loss terms, respectively.

3.2.2 Implementation details

The proposed network was trained in an end-to-end fashion to optimize the loss function; the convolution layers' weights were initialized with a zero-mean Gaussian distribution, with a standard deviation of $2/m$, where $m = f^2 \times n_f$, f is a filter size, and n_f is the number of filters; this initialization can relieve diminishing gradients and improve deeper network architectures' convergence(69).

The discriminators' learning rate γ_D was set to 10^{-5} equally for D_{HR} and D_{LR}, while the learning rate for the generators G and F was set to $\gamma_G = \gamma_D/2$, following the Two Times Update Rule (TTUR)(41), to improve GAN convergence under mild assumptions. Dropout regularization layers, applied in the generators, were initialized with the rate $p_{\text{Dropout}} = 0.8$. Leaky ReLU layers were initialized with the negative slope coefficient $\alpha = 0.1$. The loss weights λ_1, λ_2, and λ_3 were set to 1, 0.5 and 0.00001, respectively.

The training used the Adam optimizer with exponential decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.9$ during 100 epochs with batches of 16 images. On average, the training took 9-11 hours per iteration, using TensorFlow (version 2.3.0) on a shared HPC workspace with an Nvidia Tesla P100 Graphics Processing Unit (GPU). The implemtned code is available under the GNU license on https://github.com/erickcfarias/SR-CIRCLE-GAN.

Model evaluation and comparisons

In order to evaluate the trained model, conventional quantitative metrics—namely, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM)—were calculated on 1,000 CT images held out for performance evaluation. As a baseline for comparison, we also resampled the images using a Bicubic interpolation method.

3.2.3 Benchmarks

Aiming to test the effectiveness of the proposed framework, comparisons were performed against the bicubic interpolation, presented in chapter 2, which is the most used interpolation method in SR studies, and state-of-the-art methods for single image super-resolution, namely:

3.2.3.1 Image Super-Resolution Network with an Expectation-Maximization Attention Mechanism (EMASR)

EMASRN is a CNN-based model proposed to reduce model complexity by decreasing the parameter search space, while maintaining state-of-the-art performance. In order to achieve this, Zhu *et al.* (*20*) implemented an intricate architecture composed of five foundational blocks:

The **Initial Feature Extraction Block** (IFE) consists of two convolutional layers with a 3×3 kernel used to extract shallow features, which are passed as input to a **Deep Feature Extraction Block** (DFEB). This block consists of three other modules: a **Deep Projection Block** (DPB), a **Progressive Multi-Scale Feature Extraction Block** (PMSFE) and the **Expectation-Maximization Attention Block** (EMAB).

In order to perform back projection, the DPB applies three iterations of upsampling and downsampling, connected by a convolution with 1×1 kernel.

The PMSFE is implemented to improve the model's performance by extracting multi-scale features. This block performs dillated convolutions with different rates (6, 12 and 18) to the input, in order to obtain different scale features. The output of each dillated convolution is then concatenated with the lower adjacent scale feature and passed to a convolution with a 1×1 kernel and then a batch normalization and rectified linear unit activation are performed. In parallel, a pooling operation followed by a upsampling is also performed on the input, and the output is concatenated with the different scale features. Finally, a convolution is performed in order to obtain the PMSFE output.

The EMAB implements an expectation-maximization attention mechanism, which captures long range pixel dependencies on the feature map, reflecting the image's internal information with more integrity. In this block, the input is upsampled and the Expecation-Maximization is performed T times. The output is then downsampled, a rectified linear unit activation is performed and a convolution operation with a 1×1 kernel is applied.

In the reconstruction block, the output of the DFEB is subjected to a deconvolution and convolution operation and added to the upsampled low-resolution input in order to obtain the final super-resolved image.

In order to train this model as a benchmark for this research, we relied on the implementation available at `https://github.com/xyzhu1/EMASRN`. The network was optimized for ℓ_1-norm loss during 1000 epochs with $T = 4$, a batch size of 16, and a learning rate of 10^{-5} halved every 200 epochs.

3.2.3.2 Enhanced Deep Super-Resolution

EDSR was proposed by Lim *et al.* (*70*) as an enhanced version of the SRResNet (*29*) with demonstrated state-of-the-art performance and lower computational complexity. This was achieved by simply removing the batch normalization step from the residual block architecture. The batch normalization step is responsible for normalizing the feature map, which also reduces the network range flexbility and results in more blurry images. This adjustment, compared to the original version, enable to reduce memory usage in 40%, which allowed the authors to build a larger model, with 32 residual blocks.

For the EDSR model training, we relied on the SRResNet implementation available at `https://github.com/twtygqyy/pytorch-SRResNet`. The adjustments decribed in the EDSR paper were applied in this implementation and the network was trained with the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, optimizing for ℓ_1-norm loss during 500 epochs, a batch size of 16, and a learning rate of 10^{-5} halved every 100 epochs.

3.2.3.3 Cascading Residual Network

Ahn *et al.* (*71*) proposed the implementation of a cascading mechanism upon a residual network. In comparison to the SRResNet (*29*), residual blocks are replaced by cascading blocks. Each cascading block is composed of three efficient residual blocks (*71*) followed by a 1×1 convolution. The output of the cascading block's intermediate layers are cascaded in to further layers: the cascading block input and the output of each efficient residual block are cascaded into further 1×1 convolutional layers within the block. So, for instance, the input for the last convolutional layer in the cascading block would be a concatenation of: (1) the cascading block input, (2) the output of the previous convolutional layer (local cascading connection) and (3) the outputs of the first and second efficient residual blocks (global cascading connection). The authors demonstrated that this architecture was able to achieve superior performance with lower computational complexity.

In order to train this model as a benchmark for this research, we relied on the implementation available at `https://github.com/nmhkahn/CARN-pytorch`. The network was optimized for ℓ_1-norm loss, trained with the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, during 500 epochs, a batch size of 16, and a learning rate of 10^{-5} halved every 100 epochs.

3.2.3.4 Super-Resolution based on Dictionary Learning and Sparse Representation

Jiang *et al.* (26)) proposed a super-resolution learning-based method to establish a relationship between LR-HR patches through dictionary learning and sparse representation (DLSR), modelling the problem according to the following equation:

$$LR = S \times H \times HR \tag{3.2}$$

where S is the sampling matrix and H is the sparse matrix. This method assumes that the $LR - HR$ pairs of images share sparse representation coefficients, as follows (26):

$$x \approx D_h \alpha \text{ for some } \alpha \in \mathbb{R}^K \text{ with } \alpha_0 << K \tag{3.3}$$

The small patches x segmented from an HR image can be sparsely represented by a dictionary and the sparse representation coefficient α that is obtained jointly for each patch x for the input image LR in the training process.

For the DLSR model, we trained the D_l, D_h dictionaries with a size of 2048 atoms, using $100,000$ randomly sampled patches, a sparsity regularization parameter $\lambda = 0.4$ and 5×5-pixel patches with an overlap of 4 pixels between adjacent patches, as used in the original work. We varied the upscale rate to generate the 2× and 4× versions for all the tested models.

Besides these performance benchmarks, in order to further assess the performance of the proposed GAN-CIRCLE-based SR method at 4× SR, we also compared the native 4× GAN-CIRCLE SR against the sequential application of two GAN-CIRCLE instances at 2× SR, denoted as GAN-CIRCLE$^{\mathrm{x}}$.

3.3 Radiomic feature robustness analysis

The intraclass correlation coefficient (ICC) was computed to identify which features are correlated with the number of bins used during the quantization step. Given k multiple measurements to be compared (i.e., 6 different re-binnings), ICC(3,1)(62) for a two-way random-effects (or mixed effects) model was used:

$$\mathrm{ICC}(3,1) = \frac{\mathrm{MS}_R - \mathrm{MS}_E}{\mathrm{MS}_R + (k-1)\mathrm{MS}_E}, \tag{3.4}$$

where MS_R and MS_E are the mean square for rows and mean square for error, respectively.

According to the ICC values(72), we divided the features into:

- Poor robustness: ICC ≤ 0.5;

- Moderate robustness: $0.5 < \text{ICC} \leq 0.75$;
- Good robustness: $0.75 < \text{ICC} \leq 0.9$;
- Excellent robustness: $\text{ICC} > 0.9$.

We investigated how the robustness of the textural features (in terms of ICC) varies according to the different groups of images. For each group, with the aim of identifying the most robust features, the ICC was calculated by varying the number of bins considered $\{8, 16, 32, 64, 128, 256\}$. By doing so, we determined the number of robust features by varying the number of bins in the quantization step. After determining the features showing excellent robustness, we aimed to identify the most relevant features for the analysis at hand; for this purpose, we used in an agnostic way the most best known technique of dimensionality reduction: the PCA(73). For this purpose, we had to select a specific quantization setting binning; therefore, the different number of bins were perturbed, *via* mathematical morphology operations, to select the most robust setting. With more details, the original ROIs were perturbed using morphological operators (opening and closing with a 3D spherical structuring element of 1-pixel radius). Accordingly, we produced three versions for each ROI (i.e., original, opening, and closing). This procedure simulates ROI variations through consideration of intra-/inter-reader dependence during manual contouring(74). The optimal number of bins was selected after the ROI perturbation process, by considering the re-binning with the highest number of robust features. It is worth noting that the optimal binning was selected on the Original images and not on the super-resolved ones, thus adopting the most conservative choice for fair comparisons.

With the goal of carefully analyzing these variations in terms of ICC, and after the selection of the optimal re-binning setting, we assessed the importance of these features by means of a ranking procedure: we performed a PCA and we calculated a weighted average of the features extracted from the Original images, according to the first three Principal Components (PCs), to assess their relative importance. In particular, we calculated the correlation matrix (as well as the eigenvectors and eigenvalues of the correlation matrix) to identify the PCs. PCs represent the directions of the data that explain a maximum amount of variance, i.e., the directions that capture most of the relevant and non-redundant information in the data. Then, to determine the relative importance of the features for the PCs considered, we used a quadrature sum for the individual features related to the different PCs. In this way, we determined a ranking of the features by the study of their relative weights in the main components considered.

Radiomic feature extraction

The radiomic features considered in this study were computed using PyRadiomics (version 2.2.0)(75), an open-source Python package widely used for this purpose. Since this software requires image input to be in the Neuroimaging Informatics Technology Initiative (NIfTI) format(76), a preliminary step was performed to convert the original Digital Imaging and Communications in Medicine (DICOM) scan and segmentation files to this format using custom software written in MATLAB (The Mathworks Inc., Natick, MA, USA) version R2019b.

Excluding the shape-based features and first-order features (since they are independent of the rebinning), 75 3D radiomic texture features were calculated without any image filters applied from the following categories: Gray-Level Co-occurrence Matrix features (GLCM)(56, 77, 78) (24), Gray-Level Dependence Matrix (GLDM)(79) (14), Gray-Level Run Length Matrix (GLRLM)(80) (16), Gray-Level Size Zone Matrix (GLSZM)(81) (16) and Neighboring Gray-Tone Difference Matrix Features (NGTDM)(82) (5).

The radiomic features were extracted from the NSCLC radiomics CT dataset by using different quantization configurations: the number of bins varied in $\{8, 16, 32, 64, 128, 256\}$. By relying upon the slice thickness, which is the same for all CT scans included in this homogeneous subset of the whole NSCLC dataset, 3D feature computation without any resampling was used to avoid interpolation artifacts.

Chapter

4

Results and Discussion

4.1 Image super-resolution results

Fig. 4.2 shows an example of 4× super-resolved images by the GAN-SR and Bicubic interpolation, along with their PSNR/SSIM; the sample slices are randomly selected from the TCIA NSCLC CT dataset. The PSNR values were 20.99 ± 4.937 (mean ± SD) for the Bicubic interpolation, and 21.118 ± 4.828 for the GAN-SR; no statistically significant differences between the average values emerged from the one-sided Welch's adjusted test ($p = 0.15$). The SSIM values were 0.548 ± 0.234 (mean ± SD) for the Bicubic interpolation and 0.572 ± 0.225 for the GAN-based SR. The average SSIM was significantly higher for the GAN-based SR when compared through a one-sided Welch's adjusted test ($p < 0.0001$).

Although PSNR/SSIM are widely adopted evaluation metrics, some studies(18, 29) demonstrated their limitations on medical image SR tasks since images with low perceptual quality could exhibit high PSNR/SSIM values. Whereas the CIRCLE-GAN and Bicubic interpolation baseline did not show statistically different PSNR values, the GAN-generated images were less blurry with better texture, sharper edges, and visually more similar to the ground truth, as shown in Fig. 4.1.

Perceptual quality, however, does not necessarily in- crease with higher PSNR. As such, different methods, and in particular, objective functions, have been developed to in- crease perceptual quality. In particular, methods that yield high PSNR result in blurring of details. More recently, some researchers have started to use the L1 norm since models trained using L1 loss seem to perform better in PSNR evaluation. The L2 norm (as well as pixel-wise average distances in gen- eral) between SR and HR images has been heavily criticized for not correlating well with human-observed image quality [(31)

Fig. **??** shows an example of both 2× and 4× super-resolved images obtained by the considered methods. This example provides a qualitative visual assessment of

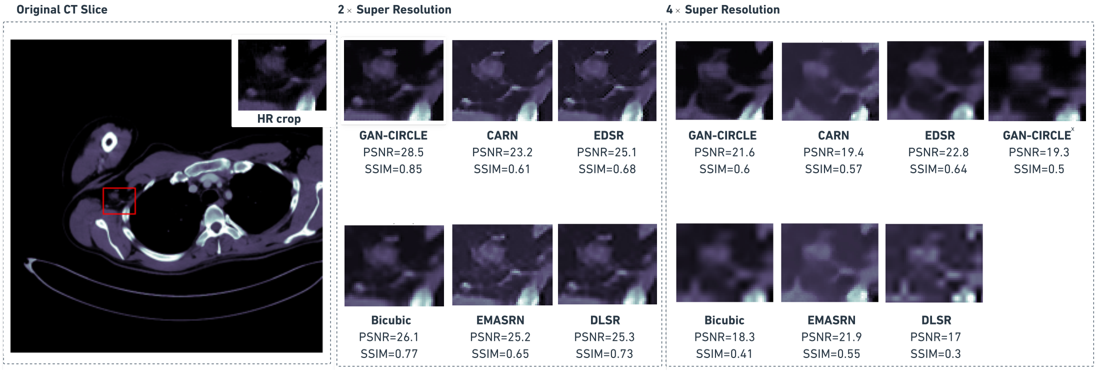

Figure 4.1: **Perceptual quality comparison** on the DeepLesion test images 2× super-resolved held out for performance evaluation, using the GAN-SR and the Bicubic interpolation method. The PSNR and SSIM values are shown at the bottom of each super-resolved image.

the super-resolved images. Fig. 4.2 reports the boxplots of the PSNR/SSIM metrics for 1,000 CT images. From the analysis of Fig. 4.2, one can see that, at 2× SR, the proposed GAN-CIRCLE-based method achieved higher median values than the other competitors for both the considered metrics (i.e., PSNR and SSIM). On the other hand, at 4× SR, the best SSIM and PSNR values were obtained with the EDSR and EMASRN SR methods. To assess the statistical significance of these results, we performed a Mann-Whitney test for pairwise comparisons (using $\alpha = 0.05$). The p-values were adjusted *via* the Benjamini-Hochberg method for multiple comparisons.

Based on the p-values yielded by the statistical test, at 2× SR, GAN-CIRCLE achieved significantly higher PSNR and SSIM values than the other competitors. The only exception is represented by the Bicubic interpolation for which the differences of the median SSIM and PSNR values were not statistically significant. At 4× SR, GAN-CIRCLE showed statistically significant differences, in terms of SSIM and PSNR, when compared against the Bicubic interpolation method and DLSR. The differences were not statistically significant when we compared GAN-CIRCLE against EDSR, EMASRN, and CARN. Finally, at 4× SR, GAN-CIRCLE^{x} produced results comparable to the ones achieved with GAN-CIRCLE.

Fig. 4.3 shows a randomly selected example from the Deeplesion dataset to endorse the quality of the produced images and assess the generalization ability of the investigated SR methods.

Although PSNR/SSIM are widely adopted evaluation metrics, some studies(*18*, *29*) have demonstrated their limitations on medical image SR tasks since images with low perceptual quality could exhibit high PSNR/SSIM values. Overall, at both 2× and 4× SR, the GAN-generated images were less blurry, with better texture, sharper edges, and visually more similar to the ground truth, as shown in Figs. **??** and 4.3.

In the downstream radiomic analyses, we focused our attention on the Original

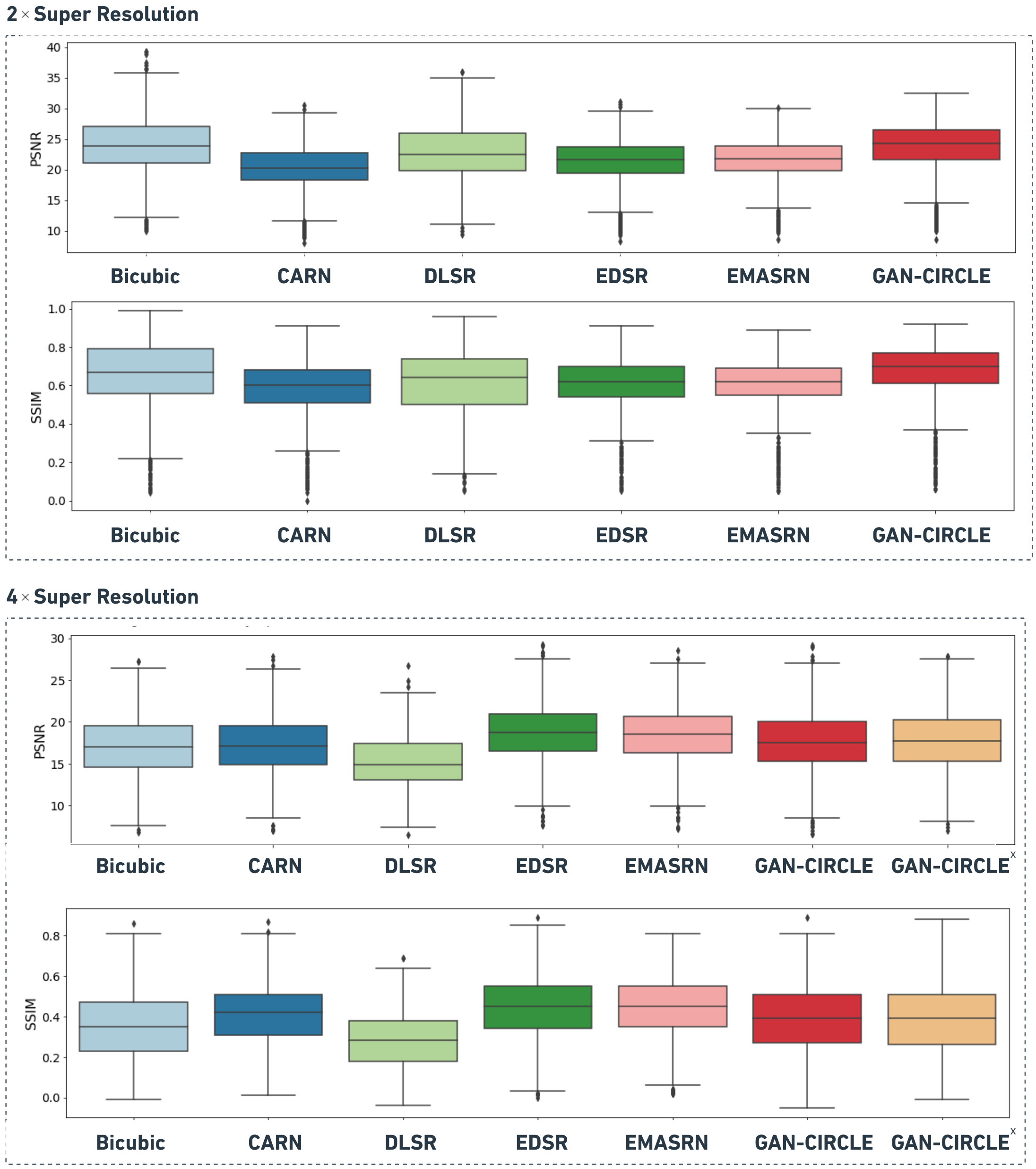

Figure 4.2: **Boxplots comparing PSNR and SSIM metrics** for 1,000 CT images held out for performance evaluation, super-resolved at 2× and 4× by using the investigated SR methods. In the case of 4× SR, GAN-CIRCLE^{x} denotes the sequential application of two GAN-CIRCLE instances at 2× SR.

images, the super-resolved images *via* the proposed GAN-SR framework (based on SPP and GAN-CIRCLE), and the Bicubic interpolation method. The Bicubic interpolation method obtained, at 2× SR, the best performance (i.e., in terms of PSNR and SSIM) among the considered SR techniques. Moreover, it is commonly available and used in medical image processing.

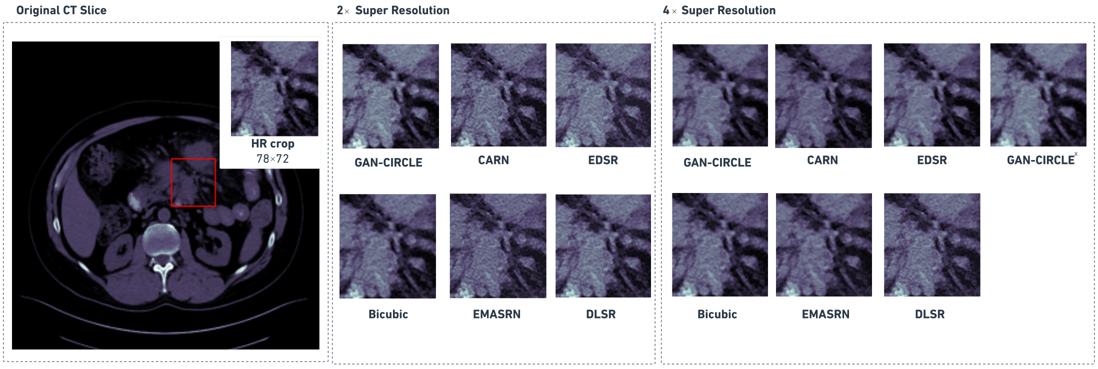

Figure 4.3: **SR example (**2× **and** 4× **factor)** using the investigated SR methods from a sample slice randomly selected from the Deeplesion dataset (held-out set). In the case of 4× SR, GAN-CIRCLE^{x} denotes the sequential application of two GAN-CIRCLE instances at 2× SR.

4.2 Robustness analysis results

In this section, we describe and discuss the results of the robustness analysis related to the textural features (in terms of ICC) according to different image groups (i.e., Original, Bicubic, and GAN-SR). Table 4.1 reports the features with excellent robustness for the considered methods. According to these values, one can observe that all the techniques taken into account produced ten features with excellent robustness. Interestingly, our GAN-SR method shows superior performance in terms of ICC for four features. Moreover, the GAN-SR technique, as well as the Bicubic interpolation, achieved moderate to good robustness for GLRLM LongRunLowGrayLevelEmphasis and GLDM DependenceEntropy, while the features extracted from the Original images resulted in excellent robustness. Table 4.2 reports the most important features according to the implemented PCA-based procedure. These four features are related to the GLCM matrix (the GLCM characterizes the texture of an image by calculating the occurrences of voxel pairs with specific values in a defined spatial relationship(*78*)) and, in particular, are the following: Correlation, IDMN, IDN, SumEntropy (Feature IDs: #1, #3, #4, #6). Of particular interest is the SumEntropy feature, defined as the sum of neighborhood intensity value differences, which showed excellent robustness with the GAN-SR method, while it showed good robustness in Original and Bicubic. Table 4.2 shows the relative difference (in terms of ICC) on the most important radiomic features between GAN-SR and the Original/Bicubic versions. With reference to the most important features, the GLCM Correlation denotes the linear dependency of gray-level values to their respective voxels in the GLCM; the Inverse Difference Moment Normalized (IDMN) is a measure of the local homogeneity of an image that normalizes the square of the difference between neighboring intensity values by dividing over the square of the total number of discrete intensity values; the Inverse

Difference Normalized (IDN) is another measure of the local homogeneity of an image that normalizes the difference between the neighboring intensity values by dividing over the total number of discrete intensity values.

Table 4.1: **Features that obtained an excellent robustness for at least of the Original, Cubic and GAN-SR image groups**.

ID	Feature name	Original	Bicubic	GAN-SR
#1	GLCM Correlation	0.980	0.979	0.984
#2	GLCM DifferenceEntropy	0.846	0.911	0.910
#3	GLCM IDMN	0.996	0.996	0.997
#4	GLCM ID	0.997	0.995	0.998
#5	GLCM MCC	0.633	0.938	0.923
#6	GLCM SumEntropy	0.822	0.897	0.905
#7	GLRLM LongRunLowGrayLevelEmphasis	0.926	0.560	0.631
#8	GLRLM LowGrayLevelRunEmphasis	0.967	0.952	0.944
#9	GLRLM ShortRunLowGrayLevelEmphasis	0.97	0.973	0.925
#10	GLDM DependenceEntropy	0.910	0.870	0.895
#11	GLDM LargeDependenceLowGrayLevelEmphasis	0.985	0.976	0.890
#12	GLDM LowGrayLevelEmphasis	0.986	0.986	0.950
#13	GLDM SmallDependenceLowGrayLevelEmphasis	0.902	0.955	0.946

Table 4.2: **Relative difference (in terms of ICC) of the GAN-SR against the Original and Bicubic versions on the most important radiomic features according to PCA analysis.**

Feature Name	Original	Bicubic	GAN-SR	GAN-SR *vs.* Original	GAN-SR *vs.* Bicubic
GLCM Correlation	0.980	0.979	0.984	0.41%	0.51%
GLCM IDMN	0.996	0.996	0.997	0.1%	0.1%
GLCM IDN	0.997	0.995	0.998	0.1%	0.3%
GLCM SumEntropy	0.822	0.897	0.905	10.1%	0.89%

According to the procedure designed for robustness in the radiomic feature, the optimal binning was found with 64 bins after the perturbation process.

In Fig. 4.4, the plots in the left column justify the use of the first three PCs, as the first three eigenvalues cover at least 85% of the trace of the covariance matrix in each group. The plots in the second column show the weights of the original features on the first three PCs, while the third column shows the relative importance of the features in the first three PCs. The most important features (in descending order), for the three groups of images, were as follows:

- Original: #1, #5, #6, #2, #10;

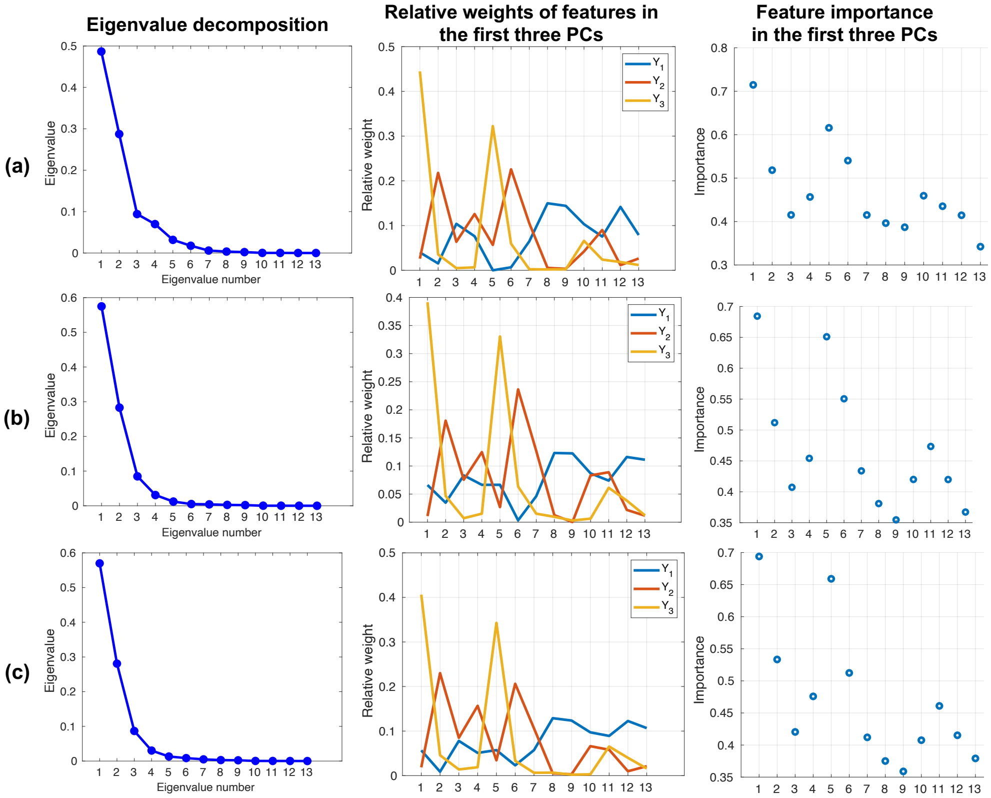

Figure 4.4: **PCA-based analysis of the importance of radiomic features for all image types: (a)** Original; **(b)** Cubic; **(c)** GAN-SR. The first column shows the line plots of the values of the eigenvalues as a function of the number of eigenvalues. This is useful for the evaluation of the PCs required. The second column shows the relative weights of the original features on the first of three PCs, while the third column depicts the relative importance of the features (according to the IDs defined in Table 4.1) in the first three PCs.

- Bicubic: #1, #5, #6, #2, #11;

- GAN-SR: #1, #5, #2, #6, #4.

Intriguingly, the features with a lower ICC in the GAN-SR method were those of less importance in terms of the PCA. Our GAN-SR method, therefore, increased the robustness of the most important features, compared to the Original and Cubic groups. These highly robust features are expected to generalize well on other and unseen imaging datasets.

4.3 Conclusion

The goal of this project was to present the first application of GAN-based image SR to radiomic studies. As a proof-of-concept, CT images were considered. In particular, the DeepLesion(63) dataset was used for training and testing the GAN-SR performance in terms of PSNR and SSIM. The performance of the proposed method was compared against the Bicubic interpolation method. Concerning the perceptual quality, experimental results demonstrated the suitability of the proposed method. With more details on the SR results obtained on the DeepLesion test images with 2× factor, GAN-SR achieved a higher SSIM value than the Bicubic interpolation method. On the other hand, while the two methods did not show statistically different PSNR values, the GAN-generated images presented better texture, sharper edges, and they looked visually more similar to the ground truth HRCT (Fig. 4.1).

In a second step, the resulting GAN-SR model was leveraged to assess the radiomic feature robustness extracted from the images of the NSCLC dataset. This assessment required the computation of the ICC to identify the most robust features against the variations of the number of bins used in the quantization step. The ICC values, calculated for the different image groups (i.e., Original, Bicubic, and GAN-SR) taken into account, showed that all the techniques obtained ten texture features with excellent robustness. Still, the proposed GAN-SR method presented superior ICC values in four of the ten features with excellent robustness. Finally, a PCA was performed to identify the relative importance of the radiomic features in the proposed GAN-SR technique. The results obtained from this analysis are particularly interesting as the features with the lowest ICC values are the ones deemed as less relevant in terms of the PCA analysis. On the contrary, GAN-SR increased the robustness of the most important features compared to the Original and Bicubic groups. The result is relevant because the highly robust features identified by GAN-SR might generalize well on other CT datasets. The results of this study could pave the way for the application of GAN-based image SR techniques for radiomics studies for robust biomarker discovery(83, 84).

Along with the novelties in lesion-focused GAN-based SR, this work belongs to the research strand dedicated to the analysis of radiomic feature robustness, with particular interest in oncological imaging. As a matter of fact, the investigation techniques used in our study were consistent with the state-of-the-art: the ICC was adopted in radiomic feature robustness analyses that assessed the impact of different imaging acquisition and reconstruction parameters(6, 7, 58), as well as image perturbations(4, 5, 8). Moreover, we identified the most important features in an agnostic manner, which is independent on a particular classification/prediction task at hand, by using a PCA-based investigation(73).

The main limitation of the proposed SR method is inherent to its lesion-focused approach, that relies on a lesion detection step for ROI identification that limits the application of this method to datasets with a pre-existent mapping of ROIs. Regarding this matter, our methodological approach could be extended to include a lesion detection task as in (*18*), to allow for CT images without lesion annotations also in the training process. Considering that our GAN-SR method currently performs only in-plane 2D image SR, to avoid the effect of slice thickness variability(*6*, *7*), GAN-based SR along the *z*-axis (i.e., yielding thinner slices) might relieve the problem related to highly anisotropic voxels(*51*, *52*). Moreover, since our GAN-SR model does not remarkably improve PSNR/SSIM values, we could conduct feature recalibration, such as *via* self-attention mechanisms, to obtain features more similar to the original images' ones(*85*–*87*). Concerning future radiomics applications, since we showed the results on a homogeneous subset of the NSCLC-Radiomics dataset, we plan to test the generalization ability of GAN-extracted radiomic features on the whole dataset, considering variations on CT image acquisition and reconstruction parameters. In particular, a classification/prediction modeling task for NSCLC staging and type would be beneficial(*64*).

Bibliography

(1) R. Gillies, P. Kinahan, and H. Hricak. "Radiomics: images are more than pictures, they are data". In: *Radiology* 278.2 (2015), pp. 563–577. DOI: 10.1148/radiol.2015151169 (cit. on pp. 1, 19).

(2) A. Zwanenburg et al. "The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping". In: *Radiology* 295.2 (2020), pp. 328–338. DOI: 10.1148/radiol.2020191145 (cit. on p. 1).

(3) I. Fornacon-Wood et al. "Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform". In: *Eur. Radiol.* (2020). DOI: 10.1007/s00330-020-06957-9 (cit. on p. 1).

(4) A. Zwanenburg et al. "Assessing robustness of radiomic features by image perturbation". In: *Sci. Rep.* 9.1 (2019), pp. 1–10. DOI: 10.1038/s41598-018-36938-4 (cit. on pp. 1, 21, 41).

(5) M. Mottola et al. "Reproducibility of CT-based radiomic features against image resampling and perturbations for tumour and healthy kidney in renal cancer patients". In: *Sci. Rep.* 11 (2021), p. 11542. DOI: 10.1038/s41598-021-90985-y (cit. on pp. 1, 21, 41).

(6) M. Shafiq-Ul-Hassan, G. Zhang, K. Latifi, et al. "Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels". In: *Med. Phys.* 44.3 (2017), pp. 1050–1062. DOI: 10.1002/mp.12123 (cit. on pp. 1, 21, 41, 42).

(7) L. Escudero Sanchez et al. "Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle". In: *Sci. Rep.* 11 (2021), p. 8262. DOI: 10.1038/s41598-021-87598-w (cit. on pp. 1, 21, 41, 42).

(8) E. P. Le et al. "Assessing robustness of carotid artery CT angiography radiomics in the identification of culprit lesions in cerebrovascular events". In: *Sci. Rep.* 11 (2021), p. 3499. DOI: 10.1038/s41598-021-82760-w (cit. on pp. 1, 21, 41).

(9) V. Sandfort et al. "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks". In: *Sci. Rep.* 9 (2019), p. 16884. DOI: 10.1038/s41598-019-52737-x (cit. on p. 2).

(10) J.-Y. Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proc. IEEE International Conference on Computer Vision*. IEEE. 2017, pp. 2223–2232. DOI: 10.1109/ICCV.2017.244 (cit. on pp. 2, 13–15).

(11) C. Han et al. "Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection". In: *IEEE Access* 7 (2019), pp. 156966–156977. DOI: 10.1109/ACCESS.2019.2947606 (cit. on p. 2).

(12) C. You et al. "CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE)". In: *IEEE Transactions on Medical Imaging* 39.1 (Jan. 2020), pp. 188–203. ISSN: 1558-254X. DOI: 10.1109/TMI.2019.2922960 (cit. on pp. 2, 6, 10, 16, 26–28).

(13) Y. Chen et al. "Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network". In: *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2018, pp. 91–99. DOI: 10.1007/978-3-030-00928-1_11 (cit. on pp. 2, 16).

(14) H. Yu et al. "Computed tomography super-resolution using convolutional neural networks". In: *IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 3944–3948. DOI: 10.1109/ICIP.2017.8297022 (cit. on pp. 2, 16).

(15) J. Park et al. "Computed tomography super-resolution using deep convolutional neural network". In: *Phys. Med. Biol.* 63.14 (2018), p. 145011. DOI: 10.1088/1361-6560/aacdd4 (cit. on pp. 2, 16).

(16) A. S. Chaudhari et al. "Super-resolution musculoskeletal MRI using deep learning". In: *Magn. Reson. Med.* 80.5 (2018), pp. 2139–2154. DOI: 10.1002/mrm.27178 (cit. on pp. 2, 16).

(17) I. Guha et al. "Deep learning based high-resolution reconstruction of trabecular bone microstructures from low-resolution CT scans using GAN-CIRCLE". In: *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*. Vol. 11317. International Society for Optics and Photonics. 2020, 113170U. DOI: 10.1117/12.2549318 (cit. on pp. 2, 16).

(18) J. Zhu, G. Yang, and P. Lio. "How can we make GAN perform better in single medical image super-resolution? A lesion focused multi-scale approach". In: *Proc. IEEE 16th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2019, pp. 1669–1673. DOI: 10.1109/ISBI.2019.8759517 (cit. on pp. 2, 16, 26, 35, 36, 42).

(19) P. Yi et al. "A progressive fusion generative adversarial network for realistic and consistent video super-resolution". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2020). DOI: 10.1109/TPAMI.2020.3042298 (cit. on p. 2).

(20) X. Zhu et al. "Lightweight Image Super-Resolution with Expectation-Maximization Attention Mechanism". In: *IEEE Transactions on Circuits and Systems for Video Technology* 8215.c (2021), pp. 1–13. ISSN: 15582205. DOI: 10.1109/TCSVT.2021.3078436 (cit. on pp. 2, 30).

(21) X. Ouyang et al. "Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond". In: *arXiv preprint arXiv:1804.02047* (2018) (cit. on p. 2).

(22) S. C. Park, M. K. Park, and M. G. Kang. "Super-resolution image reconstruction: a technical overview". In: *IEEE Signal Processing Magazine* 20.3 (May 2003), pp. 21–36. ISSN: 1558-0792. DOI: 10.1109/MSP.2003.1203207 (cit. on pp. 5, 6).

(23) A. Singh and J. Singh. "Super Resolution Applications in Modern Digital Image Processing". In: *International Journal of Computer Applications* 150.2 (Sept. 2016), pp. 6–8. ISSN: 09758887. DOI: 10.5120/ijca2016911458. URL: http://www.ijcaonline.org/archives/volume150/number2/singh-2016-ijca-911458.pdf (visited on 07/28/2022) (cit. on p. 5).

(24) E. Plenge et al. "Super-Resolution Methods in MRI: Can They Improve the Trade-off between Resolution, Signal-to-Noise Ratio, and Acquisition Time?" In: *Magn. Resonan. Med.* 68 (2012). DOI: 10.1002/mrm.24187 (cit. on p. 6).

(25) V. H. Patil and D. S. Bormane. "Interpolation for super resolution imaging". In: *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering* (2007), pp. 483–489. DOI: 10.1007/978-1-4020-6268-1_85 (cit. on p. 6).

(26) C. Jiang et al. "Super-resolution CT Image Reconstruction Based on Dictionary Learning and Sparse Representation". In: *Scientific Reports 2018 8:1* 8.1 (June 2018), pp. 1–10. ISSN: 2045-2322. DOI: 10.1038/s41598-018-27261-z. URL: https://www.nature.com/articles/s41598-018-27261-z (cit. on pp. 7, 32).

(27) C. Dong et al. "Image Super-Resolution Using Deep Convolutional Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.2 (Feb. 2016), pp. 295–307. ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2015.2439281. URL: http://ieeexplore.ieee.org/document/7115171/ (visited on 07/28/2022) (cit. on p. 7).

(28) J. Kim, J. K. Lee, and K. M. Lee. *Accurate Image Super-Resolution Using Very Deep Convolutional Networks*. Tech. rep. arXiv:1511.04587. arXiv:1511.04587 [cs] type: article. arXiv, Nov. 2016. URL: http://arxiv.org/abs/1511.04587 (visited on 07/28/2022) (cit. on p. 7).

(29) C. Ledig et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 105–114. ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.19 (cit. on pp. 7, 31, 35, 36).

(30) W. Yang et al. "Deep Learning for Single Image Super-Resolution: A Brief Review". In: *IEEE Transactions on Multimedia* 21.12 (Dec. 2019). arXiv:1808.03344 [cs], pp. 3106–3121. ISSN: 1520-9210, 1941-0077. DOI: 10.1109/TMM.2019.2919431. URL: http://arxiv.org/abs/1808.03344 (visited on 07/28/2022) (cit. on p. 7).

(31) S. Menon et al. "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models". In: (2020). arXiv: 2003.03808. URL: http://arxiv.org/abs/2003.03808 (cit. on pp. 7, 35).

(32) L. Rundo et al. "MedGA: A novel evolutionary method for image enhancement in medical imaging systems". In: *Expert Systems with Applications* 119 (2019), pp. 387–399. ISSN: 09574174. DOI: 10.1016/j.eswa.2018.11.013. URL: https://doi.org/10.1016/j.eswa.2018.11.013 (cit. on p. 7).

(33) Z. Wang et al. "Image quality assessment: from error visibility to structural similarity". eng. In: *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society* 13.4 (Apr. 2004), pp. 600–612. ISSN: 1057-7149. DOI: 10.1109/tip.2003.819861 (cit. on pp. 7, 8).

(34) A. Horé and D. Ziou. "Image quality metrics: PSNR vs. SSIM". In: *Proceedings - International Conference on Pattern Recognition* (2010), pp. 2366–2369. ISSN: 10514651. DOI: 10.1109/ICPR.2010.579 (cit. on p. 8).

(35) I. H. Sarker. "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions". In: *SN Computer Science* 2.6 (2021), p. 420. ISSN: 2662-995X, 2661-8907. DOI: 10.1007/s42979-021-00815-1. URL: https://link.springer.com/10.1007/s42979-021-00815-1 (cit. on p. 9).

(36) K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 08936080. DOI: 10.1016/0893-6080(89)90020-8. URL: https://linkinghub.elsevier.com/retrieve/pii/0893608089900208 (cit. on p. 9).

(*37*) L. Lu et al. "Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1285–1298. ISSN: 0278-0062. DOI: 10.1109/TMI.2016.2528162. arXiv: 1602.03409 (cit. on p. 10).

(*38*) C. Han et al. "GAN-based Multiple Adjacent Brain MRI Slice Reconstruction for Unsupervised Alzheimer's Disease Diagnosis". In: (2019), pp. 1–7. arXiv: 1906.06114. URL: http://arxiv.org/abs/1906.06114 (cit. on p. 10).

(*39*) S. Miao, Z. J. Wang, and R. Liao. "A CNN Regression Approach for Real-Time 2D/3D Registration". In: *IEEE Transactions on Medical Imaging* 35.5 (May 2016), pp. 1352–1363. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2016.2521800. URL: http://ieeexplore.ieee.org/document/7393571/ (visited on 07/25/2022) (cit. on p. 10).

(*40*) I. J. Goodfellow et al. *Generative Adversarial Networks*. Tech. rep. arXiv:1406.2661. arXiv:1406.2661 [cs, stat] type: article. arXiv, June 2014. URL: http://arxiv.org/abs/1406.2661 (visited on 07/25/2022) (cit. on p. 10).

(*41*) M. Heusel et al. "GANs trained by a two time-scale update rule converge to a local Nash equilibrium". In: *Proc. 31st International Conference on Neural Information Processing Systems (NIPS)*. 2017. URL: http://arxiv.org/abs/1706.08500 (cit. on pp. 11, 29).

(*42*) S. Arora and Y. Zhang. *Do GANs actually learn the distribution? An empirical study*. Tech. rep. arXiv:1706.08224. arXiv:1706.08224 [cs] type: article. arXiv, June 2017. URL: http://arxiv.org/abs/1706.08224 (visited on 07/25/2022) (cit. on p. 11).

(*43*) A. Borji. *Pros and Cons of GAN Evaluation Measures*. Tech. rep. arXiv:1802.03446. arXiv:1802.03446 [cs] type: article. arXiv, Oct. 2018. URL: http://arxiv.org/abs/1802.03446 (visited on 07/25/2022) (cit. on p. 11).

(*44*) M. Arjovsky, S. Chintala, and L. Bottou. *Wasserstein GAN*. Tech. rep. arXiv:1701.07875. arXiv:1701.07875 [cs, stat] type: article. arXiv, Dec. 2017. URL: http://arxiv.org/abs/1701.07875 (visited on 07/25/2022) (cit. on p. 11).

(*45*) I. Gulrajani et al. *Improved Training of Wasserstein GANs*. Tech. rep. arXiv:1704.00028. arXiv:1704.00028 [cs, stat] type: article. arXiv, Dec. 2017. URL: http://arxiv.org/abs/1704.00028 (visited on 07/25/2022) (cit. on p. 12).

(*46*) P. Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. Tech. rep. arXiv:1611.07004. arXiv:1611.07004 [cs] type: article. arXiv, Nov. 2018. URL: http://arxiv.org/abs/1611.07004 (visited on 07/25/2022) (cit. on p. 12).

(47) Y. Yuan et al. *Unsupervised Image Super-Resolution using Cycle-in-Cycle Generative Adversarial Networks*. Tech. rep. arXiv:1809.00437. arXiv:1809.00437 [cs] type: article. arXiv, Sept. 2018. URL: http://arxiv.org/abs/1809.00437 (visited on 07/25/2022) (cit. on pp. 14, 15).

(48) X. Yi, E. Walia, and P. Babyn. "Generative adversarial network in medical imaging: A review". In: *Medical Image Analysis* 58 (2019), p. 101552. ISSN: 13618415. DOI: 10.1016/j.media.2019.101552. arXiv: 1809.07294 (cit. on p. 16).

(49) H. Uzunova et al. *Multi-scale GANs for Memory-efficient Generation of High Resolution Medical Images*. Tech. rep. arXiv:1907.01376. arXiv:1907.01376 [cs, eess] type: article. arXiv, July 2019. URL: http://arxiv.org/abs/1907.01376 (visited on 07/27/2022) (cit. on p. 16).

(50) D. Mahapatra, B. Bozorgtabar, and R. Garnavi. "Image super-resolution using progressive generative adversarial networks for medical image analysis". In: *Comput. Med. Imaging Graph.* 71 (2019), pp. 30–39. DOI: 10.1016/j.compmedimag.2018.10.005 (cit. on p. 16).

(51) A. Kudo et al. "Virtual thin slice: 3D conditional GAN-based super-resolution for CT slice interval". In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2019, pp. 91–100. DOI: 10.1007/978-3-030-33843-5_9 (cit. on pp. 16, 42).

(52) K. Zhang et al. "SOUP-GAN: Super-Resolution MRI Using Generative Adversarial Networks". In: *arXiv preprint arXiv:2106.02599* (2021) (cit. on pp. 16, 42).

(53) P. Lambin et al. "Radiomics: the bridge between medical imaging and personalized medicine". In: *Nat. Rev. Clin. Oncol.* 14.12 (2017), pp. 749–762. DOI: 10.1038/nrclinonc.2017.141 (cit. on p. 19).

(54) H. A. Vargas et al. "Radiogenomics of High-Grade Serous Ovarian Cancer: Multireader Multi-Institutional Study from the Cancer Genome Atlas Ovarian Cancer Imaging Research Group". eng. In: *Radiology* 285.2 (Nov. 2017), pp. 482–492. ISSN: 1527-1315. DOI: 10.1148/radiol.2017161870 (cit. on p. 19).

(55) E. Segal et al. "Decoding global gene expression programs in liver cancer by noninvasive imaging". en. In: *Nature Biotechnology* 25.6 (June 2007), pp. 675–680. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt1306. URL: http://www.nature.com/articles/nbt1306 (visited on 07/29/2022) (cit. on p. 19).

(56) R. M. Haralick, K. Shanmugam, and I. Dinstein. "Textural features for image classification". In: *IEEE Trans. Syst. Man Cybern.* SMC-3.6 (1973), pp. 610–621. DOI: 10.1109/TSMC.1973.4309314 (cit. on pp. 20, 34).

(57) A. K. Jha et al. “Repeatability and reproducibility study of radiomic features on a phantom and human cohort”. en. In: *Scientific Reports* 11.1 (Jan. 2021), p. 2055. ISSN: 2045-2322. DOI: 10.1038/s41598-021-81526-8. URL: https://www.nature.com/articles/s41598-021-81526-8 (visited on 07/29/2022) (cit. on p. 20).

(58) M. Shafiq-ul-Hassan et al. “Voxel size and gray level normalization of CT radiomic features in lung cancer”. In: *Sci. Rep.* 8.1 (2018), pp. 1–9. DOI: 10.1038/s41598-018-28895-9 (cit. on pp. 21, 41).

(59) A. Traverso et al. “Repeatability and Reproducibility of Radiomic Features: A Systematic Review”. en. In: *International Journal of Radiation Oncology*Biology*Physics* 102.4 (Nov. 2018), pp. 1143–1158. ISSN: 03603016. DOI: 10.1016/j.ijrobp.2018.05.053. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360301618309052 (visited on 07/31/2022) (cit. on p. 22).

(60) A. Zwanenburg et al. “Assessing robustness of radiomic features by image perturbation”. en. In: *Scientific Reports* 9.1 (Dec. 2019), p. 614. ISSN: 2045-2322. DOI: 10.1038/s41598-018-36938-4. URL: http://www.nature.com/articles/s41598-018-36938-4 (visited on 07/29/2022) (cit. on p. 22).

(61) S. D. Walter, M. Eliasziw, and A. Donner. “Sample size and optimal designs for reliability studies”. eng. In: *Statistics in Medicine* 17.1 (Jan. 1998), pp. 101–110. ISSN: 0277-6715. DOI: 10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e (cit. on p. 22).

(62) P. Shrout and J. Fleiss. “Intraclass Correlations: Uses in Assessing Rater Reliability”. In: *Psychol. Bull.* 86.2 (1979), pp. 420–428. DOI: 10.1037/0033-2909.86.2.420 (cit. on pp. 23, 32).

(63) K. Yan et al. “DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning”. In: *J. Med. Imaging* 5.03 (July 2018), p. 1. ISSN: 2329-4302. DOI: 10.1117/1.JMI.5.3.036501. (Visited on 05/30/2021) (cit. on pp. 25, 41).

(64) H. Aerts et al. “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach”. In: *Nat. Commun.* 5 (2014), p. 4006. DOI: 10.1038/ncomms5006 (cit. on pp. 25, 42).

(65) K. Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *J. Digital Imaging* 26.6 (2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7 (cit. on p. 25).

(66) Q. Lyu, H. Shan, and G. Wang. “MRI Super-Resolution with Ensemble Learning and Complementary Priors”. In: *IEEE Trans. Comput. Imaging* 6 (2020), pp. 615–624. ISSN: 2333-9403, 2334-0118, 2573-0436. DOI: 10.1109/TCI.2020.2964201. (Visited on 05/30/2021) (cit. on p. 26).

(*67*) Y. Zhou et al. "CT-based radiomics signature: a potential biomarker for preoperative prediction of early recurrence in hepatocellular carcinoma". In: *Abdom. Radiol.* 42.6 (2017), pp. 1695–1704. DOI: 10.1007/s00261-017-1072-0 (cit. on p. 27).

(*68*) K. He et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.9 (2015), pp. 1904–1916. DOI: 10.1109/TPAMI.2015.2389824 (cit. on p. 27).

(*69*) K. He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proc. IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123 (cit. on p. 29).

(*70*) B. Lim et al. "Enhanced deep residual networks for single image super-resolution". In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017, pp. 136–144. DOI: 10.1109/TCYB.2019.2952710 (cit. on p. 31).

(*71*) N. Ahn, B. Kang, and K.-A. Sohn. "Fast, accurate, and lightweight super-resolution with cascading residual network". In: *Proc. European Conference on Computer Vision (ECCV)*. 2018, pp. 252–268. DOI: 10.1007/978-3-030-01249-6_16 (cit. on p. 31).

(*72*) E. Scalco et al. "T2w-MRI signal normalization affects radiomics features reproducibility". In: *Med. Phys.* 47.4 (2020), pp. 1680–1691. DOI: 10.1002/mp.14038 (cit. on p. 32).

(*73*) I. Jolliffe. "Principal Component Analysis". In: *Encyclopedia of statistics in behavioral science* (2005). DOI: 10.1002/0470013192.bsa501 (cit. on pp. 33, 41).

(*74*) N. Sushentsev et al. "MRI-derived radiomics model for baseline prediction of prostate cancer progression on active surveillance". In: *Sci. Rep.* 11 (2021), p. 12917. DOI: 10.1038/s41598-021-92341-6 (cit. on p. 33).

(*75*) J. van Griethuysen, A. Fedorov, C. Parmar, et al. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Res.* 77.21 (2017), e104–e107. DOI: 10.1158/0008-5472.CAN-17-0339 (cit. on p. 33).

(*76*) R. Cox, J. Ashburner, H. Breman, et al. "A (Sort of) new image data format standard: NIfTI-1". In: *NeuroImage* 22 (2004) (cit. on p. 34).

(*77*) R. M. Haralick. "Statistical and structural approaches to texture". In: *Proc. IEEE* 67.5 (1979), pp. 786–804. DOI: 10.1109/PROC.1979.11328 (cit. on p. 34).

(*78*) L. Rundo et al. "HaraliCU: GPU-powered Haralick feature extraction on medical images exploiting the full dynamics of gray-scale levels". In: *Proc. International Conference on Parallel Computing Technologies (PaCT)*. Ed. by V. Malyshkin. Vol. 11657. LNCS. Cham, Switzerland: Springer International Publishing, 2019, pp. 304–318. ISBN: 978-3-030-25636-4. DOI: 978-3-030-25636-4_24 (cit. on pp. 34, 38).

(*79*) C. Sun and W. G. Wee. "Neighboring gray level dependence matrix for texture classification". In: *Comput. Vis. Graph. Image Process.* 23.3 (1983), pp. 341–352. DOI: 10.1016/0734-189X(83)90032-4 (cit. on p. 34).

(*80*) M. M. Galloway. "Texture analysis using gray level run lengths". In: *Comput. Graph. Image Process.* 4.2 (1975), pp. 172–179. ISSN: 0146-664X. DOI: 10.1016/S0146-664X(75)80008-6 (cit. on p. 34).

(*81*) G. Thibault, J. Angulo, and F. Meyer. "Advanced statistical matrices for texture characterization: application to cell classification". In: *IEEE Trans. Biomed. Eng.* 61.3 (2013), pp. 630–637. DOI: 10.1109/TBME.2013.2284600 (cit. on p. 34).

(*82*) M. Amadasun and R. King. "Textural features corresponding to textural properties". In: *IEEE Trans. Syst. Man Cybern.* 19.5 (1989), pp. 1264–1274. DOI: 10.1109/21.44046 (cit. on p. 34).

(*83*) N. Papanikolaou, C. Matos, and D. M. Koh. "How to develop a meaningful radiomic signature for clinical use in oncologic patients". In: *Cancer Imaging* 20 (2020), p. 33. DOI: 10.1186/s40644-020-00311-4 (cit. on p. 41).

(*84*) I. Castiglioni et al. "AI applications to medical images: From machine learning to deep learning". In: *Phys. Med.* 83 (2021), pp. 9–24. DOI: 10.1016/j.ejmp.2021.02.006 (cit. on p. 41).

(*85*) Y. Li et al. "Super-resolution and self-attention with generative adversarial network for improving malignancy characterization of hepatocellular carcinoma". In: *Proc. IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1556–1560. DOI: 10.1109/ISBI45749.2020.9098705 (cit. on p. 42).

(*86*) M. Li et al. "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network". In: *IEEE Trans. Med. Imaging* 39.7 (2020), pp. 2289–2301. DOI: 10.1109/TMI.2020.2968472 (cit. on p. 42).

(*87*) C. Han et al. "MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction". In: *BMC Bioinform.* 22 (2021), p. 31. DOI: 10.1186/s12859-020-03936-1 (cit. on p. 42).