2 Background Knowledge

Following the initial overview of the main objectives proposed for this dissertation project, the next chapter provides the essential theoretical foundation to understand the key concepts discussed throughout the document. The basic principles of CT, ML and Image Imputation are particularly prominent.

2.1 Basics of Medical Imaging

Over the past decades, various imaging techniques have been developed to meet diverse medical needs, such as non-invasive diagnostic methods and the pursuit of personalized and patient-specific medicine. As a result, medical imaging has become one of the primary tools in modern healthcare [2, 4].

At a high level, all developed medical imaging techniques utilize various physical phenomena to capture information resulting from the interaction between body tissues and particles, compounds or waves, depending on the intended modality [2, 4].

Currently, several imaging techniques are available and their usefulness depends on the purpose and the desired information's nature to be collected at the examination time [2, 4, 27]. Regarding their applicability, the preponderance goes to screening and diagnostic tools, such as a simple obstetrics consultation, using ultrasound technology, or a complex tumour localization process, based on CT, Positron Emission Tomography (PET) or even MRI images [2, 4, 27, 28]. To facilitate understanding, Table 1 presents a comprehensive overview of the functions, advantages, disadvantages and applications of prevalent medical imaging modalities.

Table 1: Comparison between the most commonly used medical image modalities. All the tabled information derived from [2, 4].

MODALITY	FUNCTION	ADVANTAGES	DISADVANTAGES	APPLICABILITY
Ultrasounds	Uses sound waves to generate real-time images	Non-invasive; No ionizing radiation; Portable	Operator dependent; Limited penetration in obese patients	Pregnancy; Abdominal imaging; Vascular imaging
Conventional X-ray	Produces 2D internal structures images using X-rays	Quick; Not expensive; Widely available	Limited soft tissue contrast; Ionizing radiation exposure	Fractures; Lung conditions; Dental work
CT	Provides detailed 3D images using multiple angles X-rays	High resolution for small lesions detection	Higher radiation dose compared to X-rays; May require contrast dye	Trauma; Cancer staging; Vascular imaging
PET	Detects positron-emitting radioactive substances to visualize metabolic activity	Provides functional information; Detects cellular level diseases	Limited spatial resolution; Ionizing radiation exposure	Oncology; Neurology; Cardiology
MRI	Utilizes magnetic fields and radio waves to create detailed images	Excellent soft tissue contrast; No ionizing radiation	Expensive; Long scan time; Not possible for patients with metallic implants	Neurological disorders; Musculoskeletal injuries; Soft tissue imaging

2.2 Computed Tomography (CT)

Until the creation of CT, medical imaging methods were exclusively able to offer local and two-dimensional solutions, not providing structural visualization in their continuous and three-dimensional format [2, 4]. To compensate for this insufficiency, it was created a rotating X-ray tube, capable of emitting ionizing radiation beams in multiple orientations and consequently measurable by a set of detectors, incorporated in a gantry structure [2, 4]. This new feature enabled movement in a specific direction through the patient's body to capture a sequence of images, known as slices, which form a volume when properly arranged [2, 4]. Nevertheless, the final cross-sectional images were only accessible after the sensor data had undergone specific computational algorithms [2, 4].

Typically, CT images are applied to study the tissues surrounding the examination focus, concerning disorders identification, such as neoplastic masses and acute or chronic diseases' manifestations that directly interfere with the anatomical organ structure [2, 4].

Accordingly, this imaging technique can enhance its performance by fusing with other medical image recording mechanisms, such as PET, or even by using a substance, commonly called radiocontrast, which directly influences the physical X-ray absorption process and strongly instigates the valuable distinction between tissues with an approximate physiological density [2, 4].

2.2.1 Physical Principles

As previously mentioned, CT technology was designed based on the physical process of the X-ray creation [2, 4]. The production of this ionizing radiation begins with the acceleration of electrons from a heated metallic filament, driven by an artificially generated voltage. As a result, these particles travel in a specific direction until they collide with the anode, at a region known as the focal spot. [2, 4]. After this iteration, a small percentage of energy is converted into high-energy electromagnetic radiation, while major release occurs mainly in the heat. By analysing the resultant beam spectrum, it is possible to differentiate two types of X-ray. The first one, called characteristic X-ray, happens when the internal atom electrons are ejected due to the energy generated by the collision with highly accelerated particles, causing the release of high-energy radiation after a cascade process. On the other hand, the second type results from the direction change of the bomber's electron caused by the nuclear repulsion forces. The last one is called Bremsstrahlung X-ray [2, 4].

After the X-ray beam is generated, it must interact with the areas under examination during the CT imaging process, revealing differences in absorption levels among various structures. Mathematically, as expressed in Equation 1, the transmission-absorption mechanism of these electromagnetic waves can be described by the Lambert-Beer statement, which shows an exponential decrease in the penetration probability depending on the number and type of irradiated environments [2, 4].

$$I(d) = I_0 \cdot e^{-\mu \cdot d} \Leftrightarrow \ln\left(\frac{I_0}{I}\right) = \mu \cdot d \tag{1}$$

In more detail, I_0 corresponds to the amount of signal that leaves the X-ray source, while I is the signal quanta after penetration into the patient. Along its path, the X-ray beam passes through various environments with distinct physical and chemical properties. These environments have an associated attenuation coefficient, μ, and a thickness parameter, represented by d [2, 4].

In terms of detection by the gantry sensors, this process is managed by X-ray intensifiers, which convert high-energy X-rays (around 10 keV) into lower-energy optical photons. These optical photons are then transformed into electrical signals by scintillators. Subsequently, an analog-to-digital converter processes these electrical signals, transforming them into a digital representation of the original signal [2, 4].

Finally, the generation of ionizing radiation, combined with the acquisition methodologies and the ability to rotate around the subject, enables the measurement of X-ray transmission through the patient's body, as illustrated in Figure 1 [2, 4]. Consequently, the extracted information is initially captured in a raw projection, known as a sinogram, where the x-axis represents the number of measurements taken at each gantry position and the y-axis represents the measurements recorded by each detector [2, 4]. As the raw projection does not allow an easy visual interpretation, the transition to the image domain only happens after the operation of a mathematical and computational reconstruction process, called Filter Back Projection (FBP), which fluctuates depending on the CT scan acquisition type [2, 4].

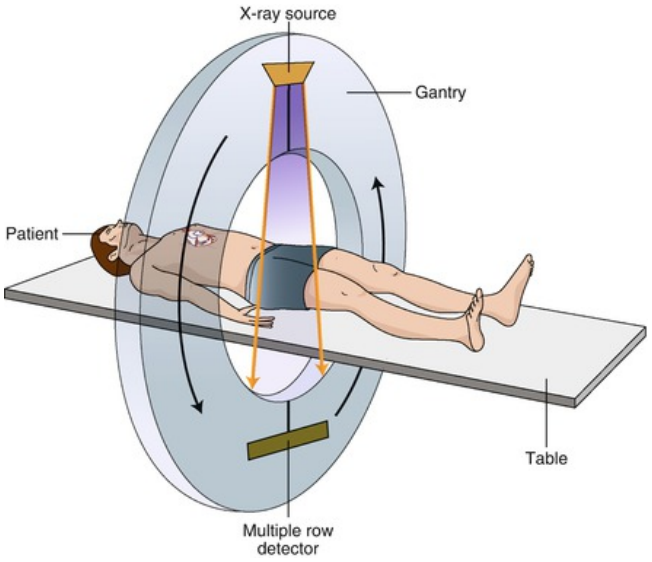

Figure 1: Schematic representation of the CT technology, from the X-ray creation to the photons detection. The source-detector system can rotate along the gantry with angle θ (adapted from [29]).

2.2.2 Clinical Applications

In the clinical field, the detailed CT three-dimensional images of the body's internal structures have solidified their importance across several medical disciplines in recent decades [2, 4]. The practicality, speed, efficiency, comprehensiveness and precision of this imaging technology have facilitated its use in complex cases, which require meticulous attention to details, as well as in urgent scenarios where rapid and accurate diagnostic imaging is essential [2, 4].

In this sense, the main areas where CT is frequently used in modern medicine include:

- **Cardiology:** CT contributes to congenital cardiac anomalies detection and cardiac intervention planning. In particular, CT provides detailed images of the heart condition during the pre-surgical phase, allowing an advanced understanding of the extent and precise location of areas needing surgical intervention [2, 4];

- **Neurology:** This imaging modality allows the assessment of cerebrovascular accidents or other intracranial haemorrhage consequences through the rapid and accurate identification of the affected brain areas. This precise uncovering helps to mitigate the risk of permanent brain damage, commonly linked with these conditions [2, 4];
- **Oncology:** The tumour masses detection in multiple body regions by CT enables a precise localization and staging of the existent carcinogenic tissue, especially when combined with PET. Additionally, it significantly impacts the treatments' evaluation, such as radiotherapy and chemotherapy, by allowing continuous adjustment of prescribed plans [2, 4, 28];
- **Orthopedics:** The three-dimensional visualization of highly complex fractures, derived from the CT, facilitates better planning of surgical procedures. It also plays a crucial role in evaluating degenerative joint diseases, aiding in informed medical decision-making [2, 4];
- **Pulmonology:** This imaging method releases a diagnosis of a wide range of lung diseases, such as interstitial diseases or lung infections. Its ability to differentiate between various lung textures under specific conditions enables precise identification of the extent and severity of potential lung injuries, offering critical insights for treatment planning [2, 4].

Since the lung is the primary focus of the current dissertation project, the role of CT in examining lung tissues is explored in greater detail through the following subtopics.

2.2.2.1 Lung-focus CT

When there is a medical need to visualize the lungs via CT, health professionals typically conduct an examination focused specifically on the thoracic region. This approach enhances the images' details and reduces the required radiation exposure compared to a full-body CT [30].

Given the unique characteristics of thoracic tissues, chest CT presents specific challenges. Since the lungs are primarily composed of air and have low density, the scans must enhance contrast to make all details within the respiratory system more visible [30]. Unlike other CT applications, this specialization requires high sensitivity to detect even small density variations between structures, allowing for the clear differentiation of high-level features within the lungs [30].

In addition, CT focusing on the thoracic region generally has a high resolution, with a small inter-slice thickness, allowing the detection of small-scale lesions in pulmonary tissue [30].

2.2.2.2 Lung Anatomy in CT Imaging

Anatomically, the lungs consist of pulmonary parenchyma — a tissue specialized for air exchange — blood vessels and an external membrane, known as the pleura [30].

Under normal conditions, when transposed to the CT image domain, the lungs exhibit a homogeneous hypodense texture with reduced opacity, resulting in their appearance in darker tones. This textural homogeneity generally indicates efficient air distribution throughout the pulmonary parenchyma [30]. At the same time, it is also possible to identify blood vessels in chest CT images, due to their higher density compared to surrounding lung tissue, causing them to appear in lighter tones [30]. These blood vessels are uniformly distributed throughout the lung, with a reduction in calibre as they extend towards more peripheral locations [30]. This gradient in vessel size explains the textural variation observed in images closer to the lung's base and apex.

Regarding the areas adjacent to the lungs, structures, such as ribs or vertebrae, have a higher density, absorbing more radiation dosages — hyperdense regions [30].

2.2.2.3 Pulmonary Diseases in CT Imaging

Knowing the standards of normal lung anatomy in advance, most algorithms designed to detect lung diseases or injuries rely significantly on identifying the textural characteristics of CT images [30]. For this reason, preserving structural and textural details in medical images revealed essential for accurate diagnosis [30].

Artefacts and imaging errors that compromise image high-level patterns can significantly reduce the accuracy and reliability of these data, thereby impacting the effectiveness of specialized models used for detection and diagnosis [30]. Thus, all the mechanisms designed to remove artefacts and address imaging errors must also consider and preserve textural patterns to effectively counteract these errors. Consequently, these mechanisms should be capable of reconstructing not only normal patterns but also potential injuries or pathologies [30].

Among the well-known patterns associated with potential lung pathologies, as illustrated in Figure 2, the following are notable:

- **Consolidation Pattern:** This pattern is marked by a uniform increase in pulmonary opacity, which leads to the obstruction of bronchial and vascular margins. It typically arises from chronic diffuse lung diseases, including chronic eosinophilic pneumonia, organizing pneumonia, lymphoma and pulmonary adenocarcinoma [30];
- **Ground-Glass Opacity Pattern:** Ground-glass opacity is characterized by pulmonary tissue's opacity increase, appearing as a lighter tone as expected in a normal situation. Additionally, this pattern does not reveal an obstruction to the visualization of underlying bronchovascular structures. Visually, this effect can be compared to ground glass. This phenomenon often occurs in cases of lung inflammation, such as pneumonia [30];

- **Honeycombing Pattern:** Honeycombing is a pattern visually similar to a honeycomb, where pulmonary cysts with well-defined walls, tending to be round or oval, are clustered in subpleural, peripheral and basal regions. This phenomenon typically appears in patients with pulmonary fibrosis [30];
- **Mosaic Attenuation:** By analogy to a mosaic, this pattern is described by the grey shade variation between adjacent lung regions, indicating disruptions in the respiratory process and blood supply. This pattern is commonly associated with conditions such as pulmonary oedema or haemorrhage [30];
- **Nodules Pattern:** Characterized by hyperdense structures with a tendency to be solid. At the CT level, due to these physical differences, they often appear as rounded masses with an opaque tone, distinguishing them from the surrounding pulmonary interstitium. While most nodules are benign and result from small lung lesions, there are instances where these nodules indicate neoplastic conditions. In this last case, it is common to distinguish two lung cancer types: Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC) [14, 30];
- **Reticular Pattern:** Reticular pattern is defined by the lungs appearing as a network of thin and interconnected light lines, giving them a mesh appearance. Such an arrangement can result from various conditions, including lung infections or medication reactions [30].

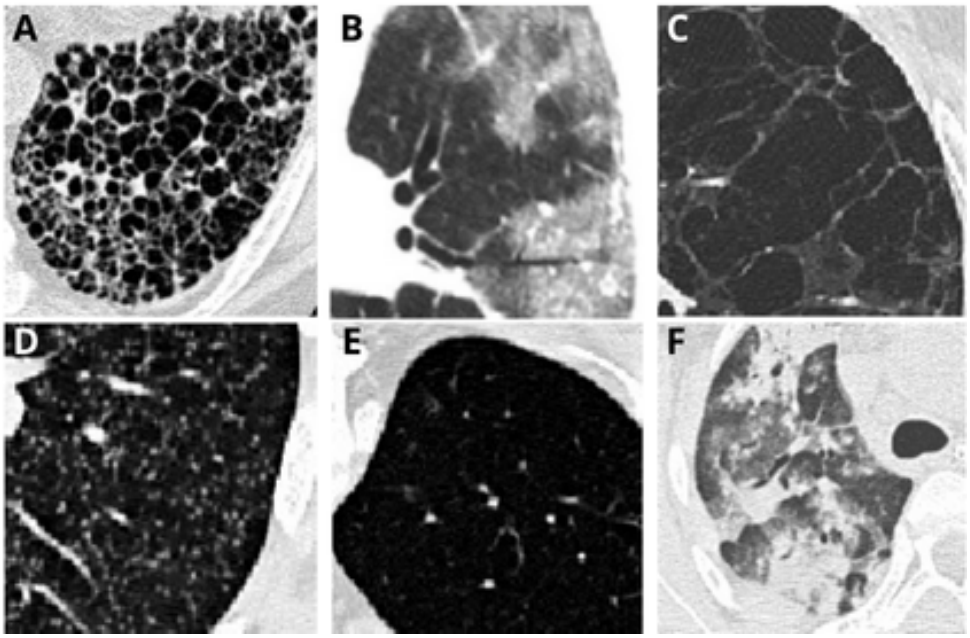

Figure 2: Examples of different lung textures potentially visualized in a CT images. These include: (A) Honeycombing Pattern; (B) Ground-Glass Pattern; (C) Mosaic Pattern; (D) Nodules Pattern; (E) Health Pattern; (F) Consolidation Pattern (adapted from [31]).

2.3 Image Errors and Artefacts

Data integrity is fundamental to ensure the accuracy, consistency and reliability of several medical decisions [32]. However, the crescent complexity of imaging techniques has led to the emergence of factors that compromise the medical data stability and create an environment prone to errors [32].

Among the various known scenarios, the transmission of information between systems stands out as one of the main error sources [32]. If not properly handled, this can introduce interoperability errors, leading to the corruption or modification of the original data, which could severely impact the diagnosis accuracy and the treatment's continuity [32]. To address these problems, standards such as Digital Imaging and Communications in Medicine (DICOM) and Neuroimaging Informatics Technology Initiative (NIfTI) were developed [32]. These standards ensure the coherence of storage, transmission and communication of all information to be shared, thereby preventing data loss during these highly complex and potentially erratic processes [32].

Another scenario that leads to collected information becoming unusable or failing the minimum quality necessary for optimal medical assessment can occur during the acquisition momentum [3]. Physical, instrumental and human factors are the primary contributors to the quality loss in the performed scans [3]. In the case of CT images, various artefacts can be observed, each with different associated causes and patterns, as discussed below:

Physical-based Artefacts;

- **Metallic non-anatomical objects:** Structures with high radiation absorbance properties, such as surgical or dental implants, create interference artefacts and dark regions in the surrounding area. Typically, the artefact reduction involves filtering procedures to alleviate visible noise [2–4]. However, managing these effects remains challenging despite these mechanisms [2–4];
- **Beam hardening and cupping:** These artefacts commonly manifest as lower material density readings than expected, particularly in the central and internal structures of tissues with significant thickness [2–4]. This phenomenon occurs due to increased X-ray absorption at the periphery and the challenges in radiation penetration through thick objects [2–4];
- **Photon starvation:** These errors occur when the amount of data acquired with the sensors is insufficient for optimal reconstruction, both physically and computationally [2–4]. Typically, this loss of information is mitigated by accurately quantifying the ideal radiation dose for each subject [2–4];

- **Scatter Artefacts:** Characterized by reduced clarity and contrast in the final image, these errors stem from the capture of low-energy scatter radiation [2–4]. This scatter radiation arises from the multiple interactions between the initial beam and the penetrated tissues, leading to the dispersion of the radiation [2–4].

Instrumental-based Artefacts:

- **Gain Artefact:** When there is an incorrect calibration of the CT scanner, it is common to appear some noise patterns, such as lines, which can severely affect the exam results [2–4];
- **Ring Artefacts:** Mostly associated with defects or poor calibration of the gantry detectors, this lack of data capture leads to the artefacts appearing in the concentric rings format on the final images [2–4];
- **Software Artefacts:** This type of error occurs when there are inaccuracies in the algorithms responsible for converting data from the sinogram domain to the image domain [2–4].

Human-based Artefacts:

- **Motion Artefacts:** These errors are caused by patient movement during acquisition. Generally, these are voluntary movements due to discomfort felt during the acquisition interval. However, errors can still occur due to involuntary movements, such as respiratory motion [2–4, 22]. The image generated from this artefact type typically presents significant visual distortions [2–4, 22];
- **Limited Field of View (FOV) Artefact:** Errors resulting from poor positioning of the patient concerning the scanner's FOV, leading to a loss of medically relevant information and the appearance of artefacts at the edges of the final images [2–4, 22];
- **Health professional-based Artefact:** Professionals who work with this type of technology must pre-program essential scanner parameters tailored to the exam type and the patient. However, artefacts frequently arise from inadequate parameter management [2–4, 22]. These errors often result from the incorrect definition of the radiation dosage administered to the patient, leading to underexposed or overexposed images. Furthermore, an erroneous application of filtering algorithms by professionals responsible for post-processing mechanisms can introduce additional errors into scan data [2–4, 22].

As previously listed, several CT artefacts exhibit interference patterns that lead to information loss and decontextualization of surrounding anatomical data, manifesting as distortions and visual gaps. These disruptions generally necessitate reconstruction processes to recover the missing data. Figure 3 provides examples of some of these cases.

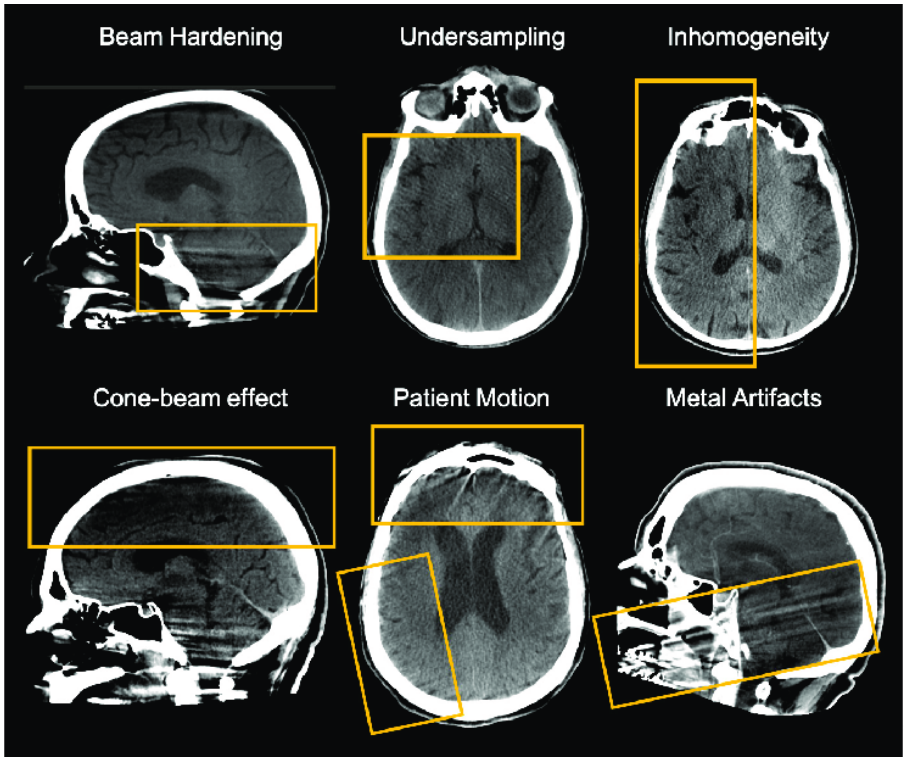

Figure 3: Examples of distortion patterns caused by several types of artefacts in CT scans (adapted from [33])

Traditionally, the most direct approach to addressing these distortions has been to re-scan the patients. However, as previously discussed, this method has several drawbacks, including increased radiation exposure for patients and higher costs for healthcare providers [4, 5]. For this reason, there is a pressing need for advanced methods to address these issues while generating a new artefact-free exam using only valid information from the erroneous scan [3].

Today, ML and CV algorithms are increasingly at the forefront of these approaches, especially because these methods operate in several directions, ranging from the reconstruction of poorly formed scans to the creation of volumetric samples from just a few two-dimensional projections, for example [6–11, 34].

2.4 Machine Learning (ML)

According to Tom Mitchell's definition, Machine Learning (ML) corresponds to the set of algorithms that have the ability "to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [35]. This self-learning power, rooted in the development of the Perceptron by Frank Rosenblatt in 1958, has facilitated the widespread adoption of these complex models across various scientific fields [36]. In essence, ML techniques involve the autonomously understanding relationships between different variables, eliminating the need for explicit programming to dictate how specific tasks should be

performed [36].

With a high-spectrum perspective, ML models can be classified into supervised, unsupervised, semi-supervised and reinforcement learning paradigms [35]. Supervised algorithms, typically employed for classification and regression tasks, utilize labelled datasets, where the ML model is trained to learn the mapping between input features and their corresponding labels [35]. In contrast, unsupervised learning models autonomously identify relationships and discover patterns within data, often clustering similar instances together [35]. Semi-supervised learning combines aspects of both supervised and unsupervised approaches. These models operate on a substantial amount of unlabeled data and a smaller set of labelled data, using the labelled examples to guide the unsupervised learning process [35]. Lastly, reinforcement learning requires direct interaction with the environment, where the model adapts internally based on a reward mechanism accomplished to its actions [35].

In recent years, ML algorithms have increasingly been utilized to process and analyze large volumes of data, aiming to mimic some of the human brain cognitive functions, such as pattern recognition and data analysis [36]. For instance, Deep Learning (DL), a specialized subset of ML, employs Neural Network (NN) with multiple layers to detect high-level patterns in unstructured data that might be challenging for traditional methods or human perception to identify [36].

Delving deeper, NN are typically composed of several layers, each containing numerous artificial neurons. These elements are characterized by their weight (w_n), bias (b) and activation function (f), as illustrated in Figure 4 [35]. In essence, each neuron computes the weighted sum of its inputs, adds a static bias value and then passes the result through an activation function to determine the final neuron output [35]. During the learning process, NN parameters are updated using backpropagation mechanisms, facilitating the convergence towards the objective function [35].

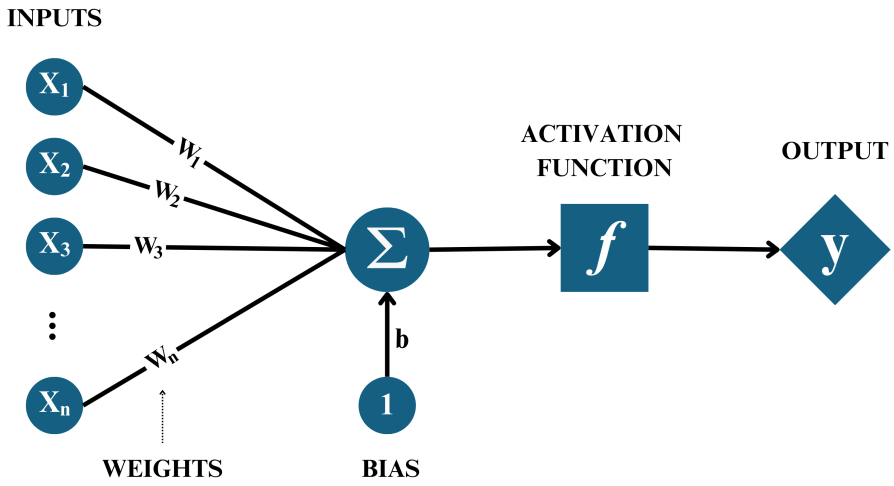

Figure 4: Artificial neuron arrangement. In the scheme, x_n, w_n, b, f and y represent the inputs, the input weights, the associated bias, the activation function and the outputs, respectively.

This learning phenomenon is commonly linked to an extensive optimization problem, where the goal is to adjust the model's outputs to align with the provided training data [35]. The perception of a good or bad fit is contingent upon the "loss function", which acts as a mathematical cost function, penalizing any model evolution that diverges from the intended direction [37].

The design of the loss function plays a pivotal role in the models' convergence, as it establishes the conditions under which the models receive penalties, thus guiding them to align with expectations outcomes. This quantification of "error", between the expected and the model-generated outputs, enables gradient calculation and the application of backpropagation techniques for NN's weights adjustment [35]. Consequently, during each epoch iteration, the model becomes increasingly accurate and fitted to the training data.

However, it is imperative to note that excessive fine-tuning may lead to a failure in model generalization, where NN exhibit high efficiency for the training dataset but poor performance for the test samples. This phenomenon, known as overfitting, is typically associated with suboptimal implementation of the ML algorithms [35].

In medicine, the applications of ML algorithms are diverse and continuously expanding in variability and complexity. Generally, these models emerge as supportive tools for healthcare professionals, operating in several contexts [38, 39]. The existing applications, superficially illustrated in Figure 5, can therefore be subdivided based on their performing medical background, according to the following order:

- **Diagnostic Tasks:** Classification models are designed to identify and categorize various pathologies or clinical conditions based on data from diverse biosignal sources [40, 41];
- **Predictive Analysis:** Unlike classification models, which are characterized by categorical labels, regression models generate continuous outputs, giving them a predictive nature and, consequently, a slightly higher level of complexity [40];
- **Intermediate Image Analysis:**
  - Segmentation: Models that automatically differentiate and delineate anatomical structures are commonly employed in medical and surgical procedure planning [40];
  - Detection: Being an extension of classification and segmentation problems, these models detect and locate features of interest, highlighting them using bounding boxes [40];
- **Advanced Image Processing:**
  - Inpainting: Models used to reconstruct corrupted or missing parts of the image. These are generally applied for noise or artefact removal [40].

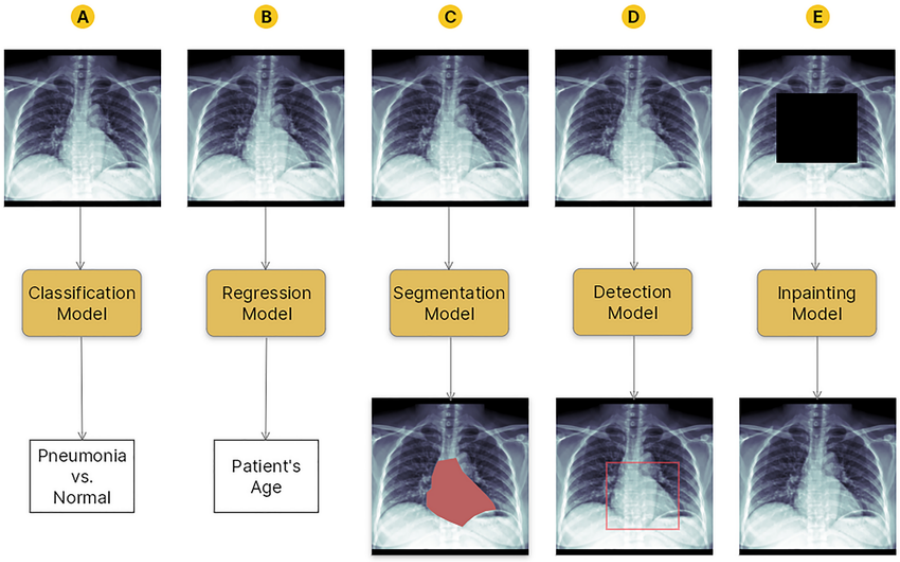

Figure 5: Illustration of some well-known ML models applied to medical imaging field: (A) Classification; (B) Regression; (C) Segmentation; (D) Detection; (E) Inpainting (adapted from [40]).

In addition to the precedent applications, with the advancement of ML models in modern medicine, other algorithms are also designed to translate between different medical image types [42], enhance medical image resolution [43] and assist in the creation of new drugs and molecular combinations [44].

2.4.1 Neural Network Architectures

In recent decades, driven by the increasing demand for ML-based solutions, the scientific community has developed several NN architectures to solve a wide range of problems. This section outlines some of the most common ones found in the literature.

2.4.1.1 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a feed-forward NN that incorporates filters, commonly known as kernels, to extract high-level features from the input data. Unlike pre-processing models, CNN autonomously learn the kernel's weights, providing them with significant flexibility during the feature extraction process [45]. These models were first proposed by Lecun et al. [46], in 1998, during the development of *LeNet-5* for handwritten digit recognition [46]. They introduced the reduction of several trainable parameters compared to prevailing networks, which typically utilized 1×1 kernels. Even so, integrating images as data sources significantly increases computational complexity due to the matrix calculations involved in CNNs. To address this challenge, Graphics Processing Unit (GPU) acceleration is commonly employed [45].

In general cases, CNNs are structured by three layer classes, as illustrated in Figure 6: Convolutional layer, Pooling layer and Fully-connected layer [45].

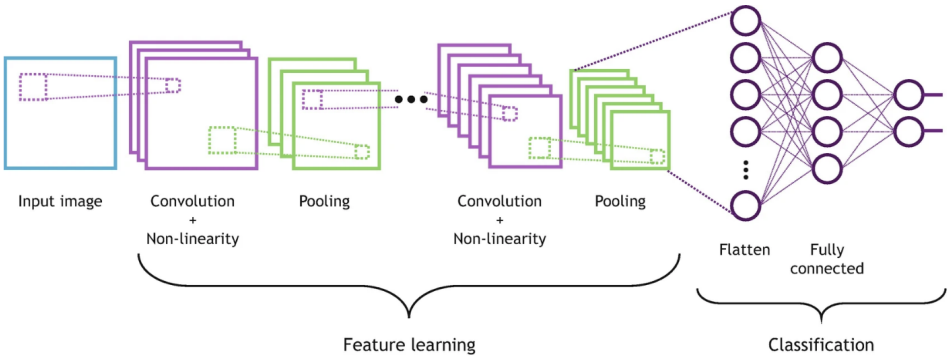

Figure 6: Classic CNN architecture. The first segment uses convolution operations for feature learning, while the second segment handles the classification or regression task (adapted from [45]).

The first layer mentioned, the Convolutional layer, aims to perform convolutional operations to create a feature map from the input data. This feature map can be visualized as a "volume" in the feature space, which ideally should fully describe the original data. The output size of these layers varies based on chosen hyperparameters, including stride, padding, dilation and kernel size [45].

In the Pooling layer, the main function is to perform a downsampling step, where the input layer data size is reduced. Similar to the Convolutional layer, it utilizes a kernel, but instead without trainable weights. This layer employs aggregation functions like "max pooling", which selects the maximum value pixel in the receptive field to send to the output array, or "average pooling", which computes the average of pixel values over the receptive field as the filter moves [45].

Finally and especially for classification problems, where a final vector associated with a particular class is required, Fully-connected layers are employed to facilitate the decision-making process based on the feature extraction performed by the preceding layers [45].

Within this family of NN, transposed convolutions are often employed, especially in networks like U-net and Fully Convolutional Network (FCN) [45]. In these applications, the convolution process is effectively reversed to expand feature map information. This configuration is characterized by a contracting (or encoder) path followed by an expanding (or decoder) path [45]. However, it's important to note that due to information loss during the encoding phase, the reconstruction process is not fully reversible [45]. Despite this limitation, transposed convolutions have successfully extended basic convolutional algorithms to handle more complex image-to-image tasks. This advancement was demonstrated by Ronneberger et al. [47], in 2015, with the introduction of the first automatic segmentation mechanism for medical images [47].

Ultimately, there is also a three-dimensional variation of CNN, known as CNN-3D. This variation enables feature extraction across multiple channels simultaneously, allowing for the identification of multidimensional patterns. CNN-3D networks are typically used for analyzing volumetric data, such as videos or imaging volumes.

2.4.1.2 Generative Adversarial Networks (GAN)

In 2014, Goodfellow et al. [48] introduced one of the most significant generative frameworks to date, known as Generative Adversarial Network (GAN) [48]. Unlike traditional models, GANs involve two NN operating simultaneously, as outlined in Figure 7. Surprisingly, instead of working together, these networks act as adversaries. The learning process evolves as each network tries to minimize its loss function while attempting to outperform the other, as one network's gain directly corresponds to the other's loss [42, 48, 49].

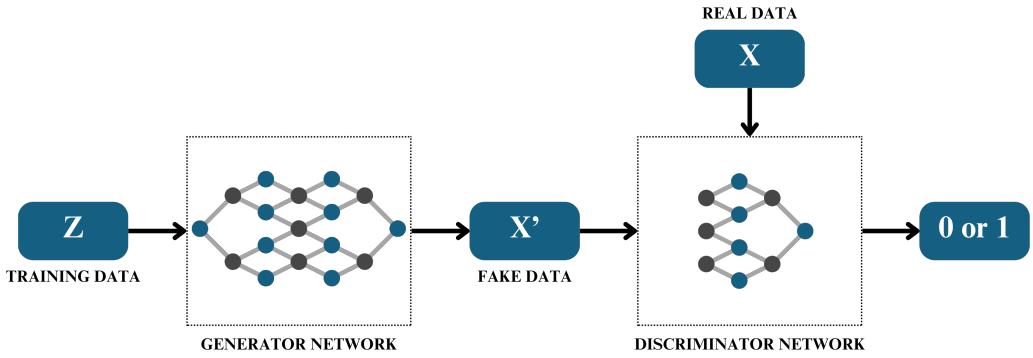

Figure 7: Basic GAN architecture. Consider Z the model input image with missing areas, X' the resultant reconstructed image and X the original image without missing information.

In basic terms, a GAN comprises a generator network, which learns to produce new data from a given input, and a discriminator network responsible for distinguishing between artificial and real data [42, 48, 49]. During training, the discriminator learns from the generator's outputs, creating a dynamic in which the generator strives to produce increasingly realistic images to maximize the error detected by the discriminator [42, 48, 49]. In the ideal scenario, after GAN training, the discriminator should be unable to differentiate between real and artificially generated data, resulting in an authenticity probability of approximately 50%, for both cases.

From an allegoric perspective, a GAN can be described as follows: "We can think of the generator as being like a counterfeiter, trying to make fake money, and the discriminator as being like police, trying to allow legitimate money and catch counterfeit money. To succeed in this game, the counterfeiter must learn to make money that is indistinguishable from genuine money, and the generator network must learn to create samples that are drawn from the same distribution as the training data" [49].

Within this family of frameworks, the Vanilla GAN is highlighted as the original model [42], while the Conditional GAN is noted for enhancing input information for both the generator and discriminator, enabling the computation of conditional probabilities rather than joint probabilities [42]. The Deep Convolutional GAN is distinguished by its generator, which is built from a complex CNN, while the Cycle GAN is recognized for its application in image-to-image translation tasks [42].

However, all these types can be summarized in a global mathematical form as a simple optimization problem. For more basic cases, such as the Vanilla GAN, this can be defined by the following general expression, according to Equation 2 [42, 48, 49].

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[log(1 - D(G(z)))] \tag{2}$$

It should be considered $p_{data}(x)$ the ground truth data distribution, $p_{data}(z)$ the artificial-generated data distribution, G the generator network and D the discriminator network [42, 48, 49]. Note that the logarithmic scale is applied to convert probability distributions generated by the discriminator network into simple sums, making the optimization process more manageable [42, 48, 49].

Through this formulation, it is also possible to understand that the generator network tries to minimize $V(D,G)$ to generate a sample indistinguishable from the real one [42, 48, 49]. Conversely, the discriminator network tries to maximize $V(D,G)$, thereby facilitating the distinction between the artificial and real samples [42, 48, 49].

On a more detailed level, the previous mathematical function can be transcribed into a computational notion by the adversarial loss function, as shown in Equation 3 [42, 48, 49]. Taking into account that the objective is to reduce loss, it is needed that $log(D(y))$ and $log(1 - D(G(z)))$ tend towards zero and consequently $D(x)$ and $1 - D(G(z))$ tend towards unity value. This condition states that the ground truth image must have a discriminator's output value equal to 1 and the image artificially generated by the inpainting methods must have a value of 0 [42, 48, 49].

$$\mathcal{L}_{\text{GAN}}(G,D) = E_x[\log(D(x))] + E_z[\log(1 - D(G(z)))] \tag{3}$$

Currently, several applications utilize GAN architectures. In this context, one notable example is their effectiveness in data augmentation, one of the most crucial pre-processing processes for enhancing ML algorithm sensitivity [50]. Unlike traditional methods such as zooming, cropping, rotating and reflecting the original dataset, generative models like GANs can create highly plausible artificial examples that significantly differ from the ground truth data [50].

2.4.2 Pre-trained Architectures

With the extensive applicability of ML in society, several large companies and institutions have made substantial investments in this science field over the last few decades. As a result, several well-known networks were created, which continue to serve as the foundation for a wide range of scientific research to this day [51]. These pre-trained networks, built on massive datasets, are typically used for recognition tasks, due to their high capacity for recognizing multiple and complex patterns. This

makes them valuable tools for feature extraction and excellent starting points for specialized training (transfer learning), reducing the computational power and dataset size requirements. [51].

2.4.2.1 *VGG*

Developed in 2014 by researchers at the *University of Oxford*, this family of networks stood out for its architectural simplicity and effectiveness in classification and recognition tasks [51, 52]. The *VGG* networks are distinguished by their emphasis on network depth while employing simple and uniform convolutional layers, with 3×3 kernels, and pooling layers, used every two or three layers to reduce dimensionality and preserve the most important features [51, 52]. Notable examples include *VGG-16* and *VGG-19*, where the names indicate the total number of layers in each algorithm, as exactly demonstrated in Figure 8. Each network was pre-trained using the *ImageNet* dataset.

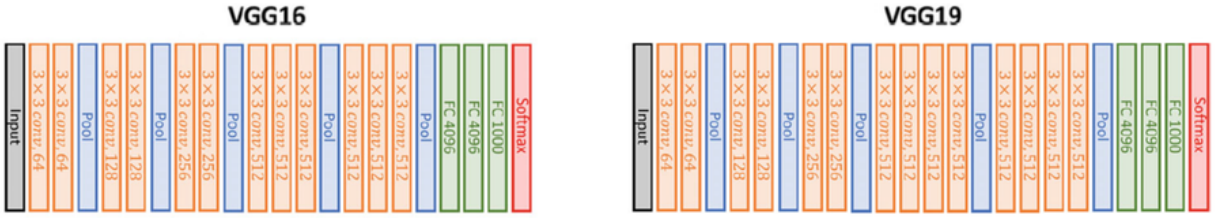

Figure 8: Comparation between *VGG-16* and *VGG-19* pre-trained architetures (adapted from [53]).

2.4.2.2 *ResNet*

Also in 2016, *Microsoft* developed a family of architectures, characterized by residual connections that facilitate information flow between layers. This innovation significantly improved network training by mitigating the vanishing and exploding gradient problems [51, 54]. Among the various models developed, the *ResNet-18* architecture, shown in Figure 9, stands out as a prominent example. This model was also pre-trained using the *ImageNet* dataset [51, 54].

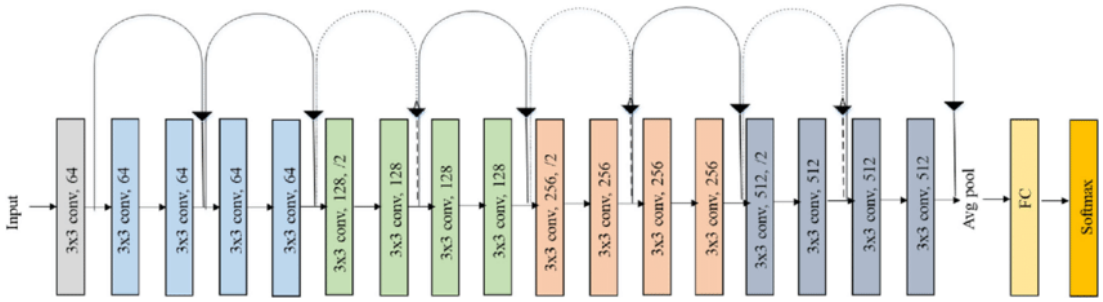

Figure 9: *ResNet-18* architecture (adapted from [55]).

2.4.2.3 *Inception*

In 2015, *Google* introduced the *Inception* networks, notable for their innovative block design that captures multiple scales simultaneously using convolution kernels of various dimensions [51, 56]. This approach enabled a compact network construction, avoiding the depth increase that would occur if each convolution were performed separately [51, 56]. Additionally, these architectures incorporate several technical advancements, such as reducing the number of operations with 1×1 convolutions and factorizing larger operations, as depicted in Figure 10. Among the different models in this family, *Inception-V2* and *Inception-V3* are the most well-known and were also trained using the *ImageNet* dataset [51, 56].

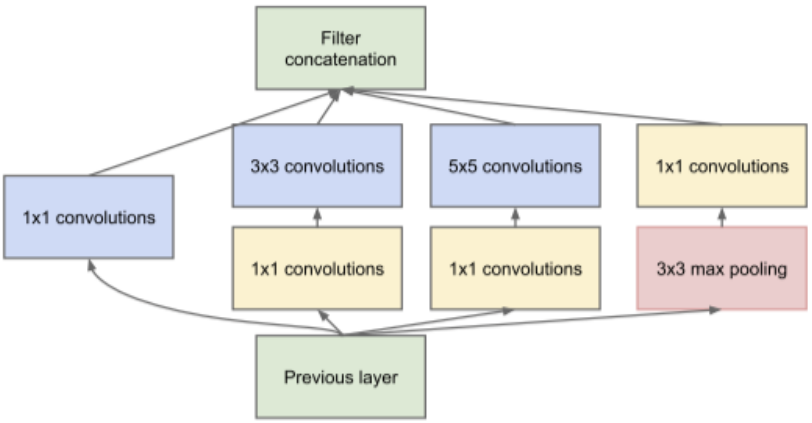

Figure 10: Esquematic representation of the *Inception* module with larger convolutions factorization capacity (adapted from [56]).

2.5 Image Imputation, Inpainting or Completion

In recent years, the integration of ML and CV algorithms in traditional image processing techniques has grown substantially, owing to the superior generalization and interpretation capabilities of these new computational approaches [6–11]. Concepts such as imputation, inpainting and completion have also gained prominence, particularly because their incorporation enhances the quality and integrity of imaging data that may be unreadable, unusable or inaccessible [6–11].

Although these three concepts often converge in imaging contexts, they are generally defined with slight differences: completion is the broadest term, referring to the inference of missing data of any type; imputation typically relates to the estimation of missing values in tabular data; and inpainting specifically applies to the imaging data [6–11].

Ultimately, due to the imaging context of the actual dissertation, these concepts — imputation, inpainting and completion — can be treated as equivalents and unified as a set of techniques aimed

at reconstructing distorted or missing regions within an image, generating high-quality structures consistent with the original input [6–11]. In essence, they encompass approaches that utilize structural and semantic information from valid data of flawed images to infer unknown or erratic details, thereby recreating data without distortions [6–11].

The initial conceptualization of these methods was introduced by Bertalmio et al. [57], in 2000, through a study that aimed to emulate fundamental image restoration techniques used by professional art restorers with a digital reconstruction algorithm, minimizing human intervention [57]. Even though, the practical application of this concept dates back to 1993 [58], when Nitzberg et al. explored a computational model for image segmentation that integrated low-level occlusion detection [58].

Both of these primordial methodologies were based on reconstructive statistical models that utilized mathematical processes of diffusion and interpolation [6–11]. These well-known sequence-based approaches are divided into two main groups based on how they fill the missing or distorted regions: diffusion-based or patch-based techniques [6–10]. In diffusion-based methods, a smooth propagation of image context occurs from the boundary towards the interior of the missing region [59]. Conversely, patch-based methods involve searching for well-matching regions in undamaged areas of the original image or external data to fill in the holes within the input image [60]. Although these methodologies have been extensively studied, their generalization falls in recreating the high-level semantic content of images, mainly, because their perception of gaps is limited to the neighbouring data, without considering the entire available image information [6–10].

To address these challenges, research over the last decade has shifted towards developing new learning-based models capable of capturing high-level patterns, that are not visible to the naked eye [6–11]. One of the main reasons for this change is the ability of ML models to interpret input images as a whole, increasing efficiency in the reconstruction process [6–11].

As a result of these advancements, image imputation techniques have expanded into various domains, such as art restoration, video processing, where they are used to reconstruct damaged frames, satellite imaging, where they help reconstruct missing data due to cloud cover, or even the medical field, where these approaches can be used to remove local image artefacts, as illustrated in Figure 11, for example [6–11].

Despite advancements in imputation algorithms, significant challenges remain, especially in the medical imaging field [6–11]. Errors in conventional imaging may be relatively harmless, but mistakes in medical image imputation can have serious consequences. This highlights the critical need for highly accurate models and their ongoing development. However, achieving such precision naturally increases computational complexity and, consequently, computational demands [6–11].

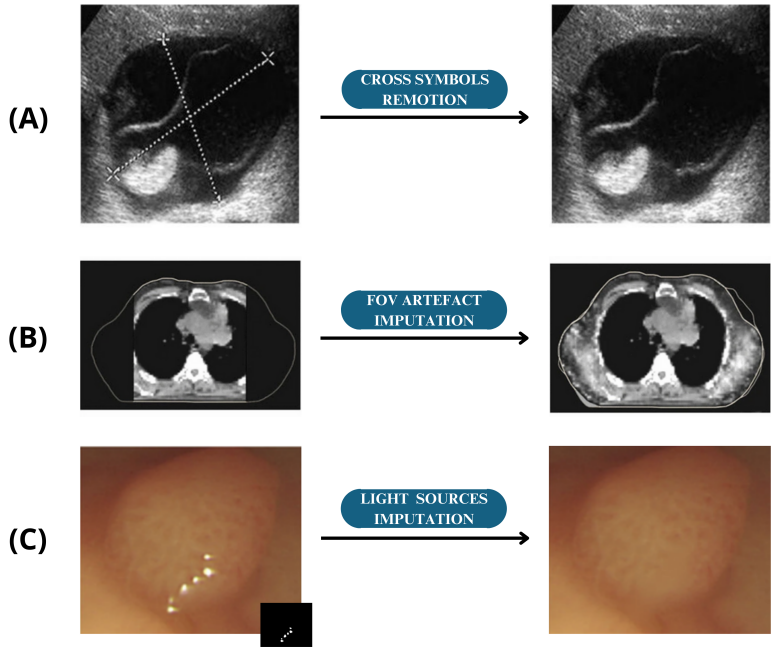

Figure 11: Examples of medical imaging artefacts correction by imputation algorithms: (A) Cross symbols imputation in ultrasound images; (B) Completion of missing areas due to FOV artefacts in chest CT images; (C) Removal of light sources in endoscopic digital images (adapted from [61–63]).

2.6 Evaluation Metrics

When using ML algorithms, it is important to understand their confidence level. As a result, the models' outputs must undergo a comprehensive evaluation to measure the discrepancies between the expected and predicted results [64–66].

Specifically for image imputation, inpainting or completion techniques, it is immediately possible to visually assess the model's reconstructive capacity. However visual assessment is highly subjective and, in many circumstances, the reconstructed patterns are extremely complex and difficult to discriminate with the unaided eye. Therefore, tangible and quantitative measures are necessary to compare different models objectively and demonstrate their real performance [64–66].

Although it is not a conventional classification in the literature, the metrics introduced will be categorized into two different groups: pixel-based and feature-based metrics.

2.6.1 Pixel-based Metrics

Pixel-based metrics focus on accuracy and precision in reconstructing missing pixels, emphasizing the intensity values of individual pixels. Typically, these evaluation metrics measure the pixel differences between the reconstructed and raw images [64–66].

2.6.1.1 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a pixel-wise measure, calculated by comparing pixels between different images. Mathematically, this error computation from paired observations is expressed in Equation 4, where X and Y are two images of dimension $m \times n$ [64, 66].

$$\text{MAE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |Y(i,j) - X(i,j)| \tag{4}$$

2.6.1.2 Mean Squared Error (MSE)

Similarly to the previous metric, Mean Squared Error (MSE) aims to calculate the squared error between corresponding pixels of the original and the model's output images, as outlined in Equation 5, where X and Y are two images of dimension $m \times n$ [64, 66],.

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (Y(i,j) - X(i,j))^2 \tag{5}$$

MAE and MSE only analyze differences between pixels. Consequently, they do not necessarily correlate with the effective reconstructive ability of inpainting models [64, 66]. Additionally, these metrics can be highly manipulable, as they yield significantly different values depending on whether the calculation is performed exclusively on the reconstructed areas or the entire image, where all regions other than the reconstructed ones are exact copies of the input image.

2.6.1.3 Peak Signal-to-Noise Ratio (PSNR)

Usually employed to evaluate the quality of reconstructed compressed images (codecs), the Peak Signal-to-Noise Ratio (PSNR) metric, as indicated in Equation 6, compares the maximum possible power of a signal (typically 255 for 8 bits image) to the power of corrupting noise that affects its accurate representation, specifically using the MSE value, computed based on the Equation 5 [64, 66]. Typically, this measurement is expressed in decibels and a higher value indicates better reconstruction capacity for the algorithm under analysis [64, 66].

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) = 20 \cdot \log_{10} \left(\frac{\text{MAX}}{\sqrt{\text{MSE}}} \right) \tag{6}$$

2.6.1.4 Structural Similarity Index Measure (SSIM)

Unlike the three previous metrics, which solely quantify errors between the reference and output model images, Structural Similarity Index Measure (SSIM) seeks to mimic human perception's ability to discern coherence when distinguishing between an artificially generated or original image [64, 66]. To achieve this, SSIM is typically calculated over several $N \times N$ windows of the inputs, based on the strong interdependencies between nearby pixels in space [64, 66]. Quantitatively, SSIM values range from -1 to 1, where a value close to 1 indicates a high similarity degree between two images [64, 66]. In Equation 7, it must be considered that μ represents the pixel-intensity mean of the respective image, while σ performs the elements of the covariance matrix.

There is also a variant of this metric called Multi-Scale SSIM (MS-SSIM), which operates complementarily on scales generated by multi-downsampling, enabling the features comparison that may not be present at a visual level [64].

$$\text{SSIM} = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)} \tag{7}$$

With $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$; $L = 2^{\#\text{bits per pixel}} - 1$, $k_1 = 0.01$ and $k_2 = 0.03$ by default.

2.6.1.5 Dice Coefficient (DICE)

Unlike previous metrics, the Dice-Sørensen Coefficient (DICE) coefficient analyzes the overlap between two distinct samples and consequently provides a pixel-by-pixel assessment of the reconstructive capacity of imputation models. Commonly used in evaluating segmentation algorithms, this metric is defined by Equation 8, where X and Y represent the two images being compared [67].

$$\text{DICE} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{8}$$

2.6.1.6 Intersection over Union (IoU)

Similarly to the previous metric, the Intersection over Union (IoU) also evaluates overlap, however a different ratio is used. In the context of inpainting algorithms, this metric is described as the ratio of the number of correctly reconstructed pixels to the union of the set of pixels predicted during reconstruction (Y) and the ground truth (X), as shown in Equation 9. This metric and DICE are generally good indicators of the effective reconstruction of specific structures, such as tumours [67].

$$\text{IoU} = \frac{|X \cap Y|}{|X \cup Y|} \tag{9}$$

2.6.2 Feature-based Metrics

In contrast to pixel-based, feature-based metrics quantify differences between high-level image characteristics, such as textures, rather than individual pixels. In these cases, features are generally extracted from the generated and ground truth images using NNs, aiming to capture information similar to what human visual perception might discern. These extracted features are then compared using specific methods designed to compute differences between feature vectors [64–66].

2.6.2.1 Inception Score (IS)

The Inception Score (IS) metric differs from previous evaluation metrics by focusing on the feature space of images, which is extracted from the pre-trained *Inception-V3* network [65, 66]. It evaluates differences in probability distributions using the *Kullback–Leibler* (KL) divergence method. To maximize the IS metric, the generated images should have low entropy across various labels while also having high coverage [65, 66]. In Equation 10, $\mathbb{E}_X$ denotes the expectation over images X, D_{KL} represents the KL divergence, $p(Y|X)$ is the conditional class distribution and $p(Y)$ is the marginal class distribution [65, 66]. In summary, IS can be considered as the exponential of the average difference between the classifications of individual images and the overall mean. These classifications, in turn, are extracted using the *Inception-V3* pre-trained network [65, 66].

$$\mathrm{IS} = \exp\left(\mathbb{E}_{\mathbf{X}}\left[D_{\mathrm{KL}}(p(Y|X)\|p(Y))\right]\right) \tag{10}$$

2.6.2.2 Fréchet Inception Distance (FID)

To complement the information provided by the IS metric, Fréchet Inception Distance (FID) is used to compare the original images with those generated by algorithms, directly [65, 66]. Initially, $2,048$ activation features (feature vectors) for each image under comparison are extracted via the *Inception-V3* pre-trained network. The Fréchet distance is then employed to measure the similarity between the feature vectors of real and generated images. The effectiveness of this metric depends on the quality of the extracted features [65, 66].

In Equation 11, μ and Σ represent the feature-wise mean and covariance matrix, respectively, of the generated and original images after being processed by the previously mentioned pre-trained NN. The term $|\mu_Y - \mu_X|_2^2$ refers to the squared distance between the two mean vectors, while Tr denotes the trace of the matrix obtained from performing operations on the two covariance matrices, Σ_Y and Σ_X.

$$\mathrm{FID} = \|\mu_Y - \mu_X\|_2^2 + \mathrm{Tr}(\Sigma_Y + \Sigma_X - 2\sqrt{(\Sigma_Y\Sigma_X)}) \tag{11}$$