POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Master's Degree Thesis

Super-Resolution Image Reconstruction using a GAN-based approach: application in Dermatology

Supervisors

Prof. Massimo Salvi

Ing. Francesco Branciforti

Candidate

Domenico Ficili

March 2023

Table of Contents

List of Tables

List of Figures

Summary

Cancer is a leading cause of global death, with skin cancer being a particularly deadly subtype. Malignant melanoma, an aggressive form of skin cancer, has been increasing worldwide. Early detection through screenings is crucial for improving survival rates. Dermoscopy, a non-invasive diagnostic method using a specialized instrument called a dermatoscope, has been widely adopted by dermatologists due to its ease of use and versatility, for evaluating suspicious skin lesions and avoiding unnecessary biopsies of stable lesions. Artificial intelligence (AI) is increasingly being used in biomedical research and clinical practice for automating diagnoses and assisting physicians. The accuracy of skin disease diagnosis is often linked to the experience of the dermatologist, making AI a useful tool for supplementary opinions and screening benign lesions. With the rise of mobile devices and digital healthcare, smartphone-based cameras and dermatoscopes are improving access to care and aiding treatment and screening. Teledermatology, a subspecialty of telemedicine that uses digital images for remote dermatological consultations, is becoming a convenient way for patients to receive direct physician evaluations through photograph submissions. However, image quality can be poor due to lighting, blurriness, and focus issues, hindering accurate diagnoses. Improving image quality is crucial, as high-resolution images are critical to precise diagnoses. To address this issue, software applications that incorporate automated computer vision and image processing tools, along with patient education on proper image capture and transmission techniques, can help restore and enhance the quality of images. The present study aims to propose a Generative Adversarial Network (GAN)-based methodology for the reconstruction of high-resolution skin lesion images from low-quality counterparts. This is achieved by selecting and processing images from the open public ISIC dataset through image processing techniques to eliminate artifacts and prepare the data for input into the GAN model. Data augmentation was also applied to increase the size of the dataset. Subsequently, the final paired dataset of low-resolution and high-resolution image pairs was generated through the application of degradation and enhancement techniques, and the Real-ESRGAN architecture was employed in the training phase. The resulting model was then used to generate super-resolution output images from available

test sets, which were evaluated through the use of Image Quality Assessment (IQA) metrics (e.g., SSIM, FSIM, and NIQE) in addition to visual assessment. The dissimilarity between restoring the original, unaltered image (GT) and its low-resolution (LR) counterpart obtained through the degradation pipeline was analyzed to comprehend the potential efficacy of the degradation and restoration processes developed. Finally, to further validate the restoration performance of the trained GAN model, both the ground truth images and the synthetic restored images from the ISIC test set were input into a deep network for skin lesion classification, permitting a comparison between the original-quality images and the reconstructed, super-resolution images with regards to their ability to accurately classify skin lesions. In conclusion, the proposed approach demonstrates potential in the reconstitution of high-resolution images of skin lesions from low-quality ones, serving as a favorable starting point for further research and optimization.

Chapter 1

Introduction

1.1 Anatomy and Physiology of the Skin

The integumentary system, also known as integument, is the body's outermost layer and consists of the skin and its associated appendages, such as hair and nails. The main function of the integumentary system is to provide a physical barrier that protects the body from external threats such as bacteria, infection, and physical injury. It also helps regulate body temperature and enables the perception of touch, pressure, and pain through receptors in the skin. Other functions of the integumentary system include protecting the body from harmful ultraviolet (UV) radiation from the sun and producing vitamin D when exposed to UV light. Overall, the integumentary system plays a crucial role in maintaining the health and well-being of the human body [1].

1.1.1 Structure

The integumentary system is made up of two main components: the skin and the accessory structures, which include nails, hair, and glands.

- **Skin:** the skin is the second largest organ in the human body and the largest component of the integumentary system. It has a surface area of about 1.5-2.0 m^2 in the average adult human, and its thickness varies across different parts of the body. The skin is made up of two layers: the superficial epidermis and the deeper dermis [1, 3]:
  - **Epidermis:** the epidermis is the most superficial layer of the skin. It provides a waterproof, protective layer that serves as the first line of defense from the surrounding external environment. It is made up of four types of stratified squamous epithelial cells: *keratinocytes* which

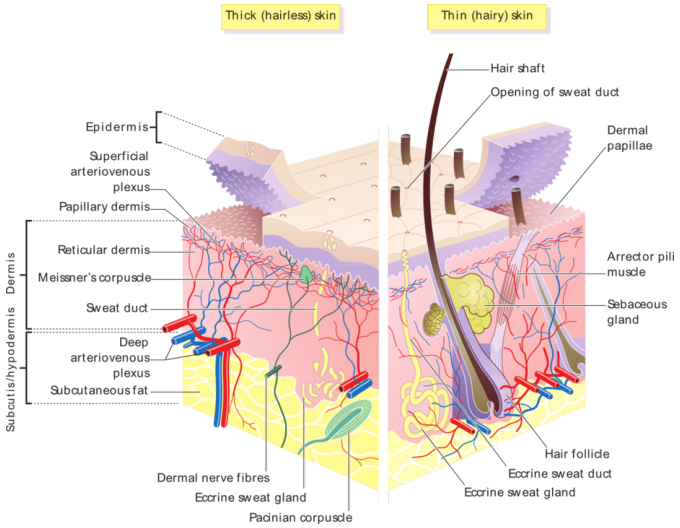


Figure 1.1: Cross-section of skin layers [2].

synthesize keratin, pigment-producing cells called *melanocytes*, *Merkel cells* involved in detecting sensation, and *Langerhans cells* which are phagocytic immunitary cells. The epidermis further divides into the following five layers, from superficial to deep:

* Stratum corneum
* Stratum lucidum
* Stratum granulosum
* Stratum spinosum
* Stratum basale

The basal layer of the epidermis contains stem cells that regenerate the epidermis as it grows towards the corneum. The epidermis receives nourishment from the dermis underneath it, as it has no blood supply of its own.

– **Dermis:** the dermis, which is immmediately deep to the epidermis, consists of connective tissues that protects the body from stress and strain while supporting the epidermis. The dermis separates into:

* **the superficial papillary layer:** it has highly vascularized, loose connective tissue which provide mechanical support for the more superficial epidermis and supplying through capillaries.
* **the deeper reticular layer:** it has an interwoven meshwork of dense connective tissue that is responsible for the toughness and elasticity of the skin.

Blood and lymphatic vessels along with nerve fibers, sweat glands and hair follicles are embedded within the connective tissue of the dermis [1, 3].

- **Hypodermis:** the hypodermis, also known as the subcutaneous tissue, is a layer beneath the skin that is primarily composed of adipose tissue. It serves several functions, including stabilizing the integumentary system, acting as an energy reserve through fat storage, and participating in thermoregulation by helping to reduce heat loss.
- **Skin appendages:** the integumentary system also includes several accessory structures, such as hair, which provides mechanical protection for the skin and helps regulate body temperature, and nails, which protect the fingers and toes. The integument also contains exocrine glands, such as sudoriferous glands, which produce sweat, and sebaceous glands, which produce oils that help keep the skin and hair moisturized. These accessory structures and glands play important roles in maintaining the health and function of the integumentary system [3].

1.1.2 Function

The integumentary system plays a crucial role in maintaining the body's homeostasis. It is composed of the skin, hair, nails, and glands and performs a variety of functions that are essential for the proper functioning of the body [1].

- **Barrier:** one of the primary functions of the integumentary system is to act as a barrier. The skin serves as the first line of defense against infections, temperature changes, and other threats to homeostasis. It is highly effective at preventing the direct entry of pathogens, protecting the body from UV rays through the secretion of melanin, and safeguarding against chemical and mechanical assaults. The skin also plays a crucial role in the wound healing process, which occurs when the body experiences trauma that causes significant damage, such as severe cuts or moderate burns [1, 4].
- **Thermoregulatory:** another important function of the integumentary system is thermoregulation. The skin being characterized by its extensive surface area, boasts a copious vasculature, enabling it to maintain and regulate thermal equilibrium through the alternate processes of vasoconstriction and vasodilation. When the body's temperature rises, blood vessels dilate to increase blood flow and maximize heat dissipation. In addition to this process, the evaporation of sweat secreted by the skin helps to cool the body.
- **Sensation:** the skin is also an important organ of sensation, as it is innervated by various sensory nerve endings that allow it to detect touch, pressure, heat,

cold, pain, and tactile sensitivity. It is often referred to as the body's largest sense organ.

- **Storage and metabolic:** the subcutaneous adipose tissue serves as a storage depot for water, fat, glucose, and vitamin D, which can be easily accessed when needed [1].

- **Endocrine**: the skin also performs an important endocrine function through the synthesis of vitamin D3.

- **Secretive:** the skin is a sophisticated secretory organ that performs various functions, such as *keratinopoiesis* and *pigmentogenesis*, which are carried out by melanocytes located in the basal layer and responsible for the production of keratin and melanin, respectively.

- **Pression homeostasis:** the highly vascularized network of the integument helps to maintain blood pressure at an appropriate level for the body's metabolic demands.

- **Immunitary:** the skin immune system (SIS) is composed of a variety of cell populations, including mast cells, dendritic cells, macrophages, lymphocytes, and keratinocytes, which are able to release a variety of mediators that can have biological effects on the skin and stimulate the entire immune system [1, 4].

1.2 Skin Cancer

Cancer represents a major cause of mortality worldwide, and skin cancer is among the malignancies with a higher occurrence rate of new diagnoses, as shown in Figure 1.2. Skin cancers are a type of malignancy that arise from the uncontrolled proliferation of skin cells. These tumours can invade and spread to other parts of the body, and there are three main types of skin cancers: melanoma, basal cell skin cancer (BCC), and squamous cell skin cancer (SCC), with the latter two being the most prevalent forms of skin cancer. It is important to recognize the early signs of skin cancer and to seek medical treatment as soon as possible to improve the chances of successful treatment [6].

1.2.1 Generalities

Skin cancer most commonly affects elderly people, even though it can afflict younger individuals as well as children on occasion. Fair-skinned persons, such as those

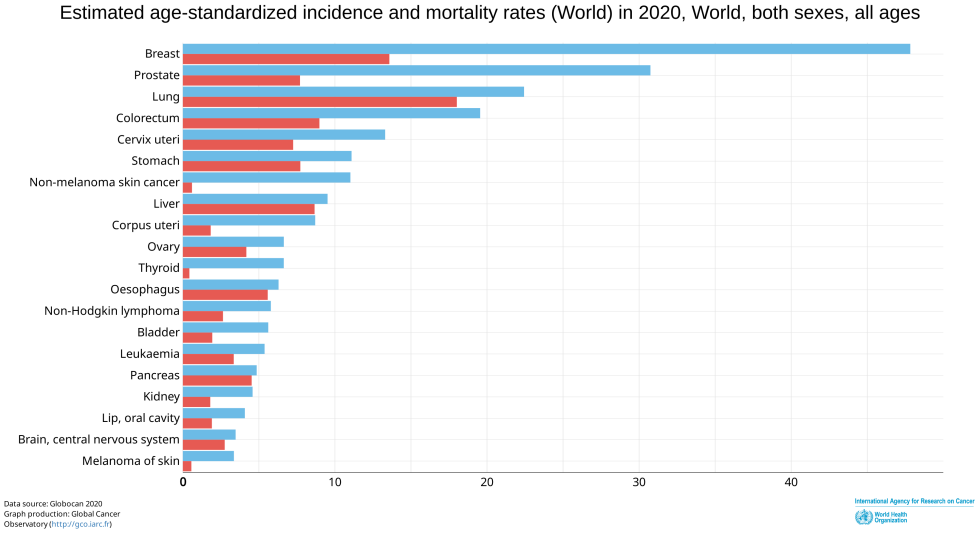


Figure 1.2: Estimated age-standardized incidence and mortality rates (World) in 2020, World, both sexes, all ages [5]

classified as Fitzpatrick [1] skin phototypes I, II, and III, are more susceptible to developing skin cancer than individuals with darker skin tones, who are still at risk. Additionally, individuals who have a family history of skin cancer or have previously had skin cancer are at an elevated risk of developing the condition [6]. Several risk factors can increase an individual's chances of developing skin cancer, including:

- Exposure to ultraviolet (UV) radiation and certain chemicals
- Ageing
- Smoking
- Drug-induced immune suppression
- Longstanding wounds
- Genetic conditions (e.g., Albinism)

Skin tumours are generally diagnosed clinically by a dermatologist, who will examine any enlarging, crusting, or bleeding lesions. A full skin examination is typically

[1]The Fitzpatrick skin phototype describes a way to classify the skin by its reaction to exposure to sunlight

performed first, and dermoscopy may be used as a preliminary diagnostic exam to detect cancers in the early stages or exclude benign lesions. In cases where a skin tumour is suspected, a partial biopsy may be taken to confirm the diagnosis. If melanoma is suspected, the patient may undergo a complete excision to establish a secure diagnosis through histopathological analysis, as partial biopsy can be misleading in some cases of melanocytic tumours [6].

1.2.2 Non-Melanoma Skin Cancer (NMSC) and other Skin Lesions

All types of skin tumours except for melanoma are referred to as non-melanoma skin cancer (NMSC). The following list highlights some of the most frequently occurring NMSC as well as other typical skin lesions that will also be encompassed within the dataset analyzed in this study:

- **Actinic Keratosis and Intraepithelial Carcinoma (AKIEC):**
  - **Actinic Keratosis (AK):** also called solar keratoses (SK), are precancerous dry scaly patches of sun-damaged skin. They are the result of abnormal skin cell development due to DNA damage caused by ageing, poor immune function and recent sun exposure [7]. They usually appear as multiple keratosis rather than solitary on sites repeatedly exposed to the sun, especially the backs of the hands and the face. Treatment from a dermatologist is very important as if left untreated, they may turn into a type of skin cancer called squamous cell carcinoma (SSC).
  - **Cutaneous Squamous Cell Carcinoma (cSCC):** also known as intraepidermal carcinoma (IEC), is one of the skin cancer common form. It presents as one or more irregular scaly plaques with often an orange-red or either brown colour. Ultraviolet radiation (UV) total exposure is the greatest risk factor as it damages the skin cell nucleic acids (DNA) and suppresses the immune response, preventing recovery from damage [8].

- **Basal Cell Carcinoma (BCC):** is by far the most common cancer in humans whose causes are multifactorials [9]. These may be triggered by exposure to ultraviolet radiation or spontaneous and inherited gene defects. BCC is a locally invasive and slowly growing nodule which presents as pink or pigmented and sometimes with spontaneous bleeding or ulceration. Fortunately they are very rarely a threat to life.

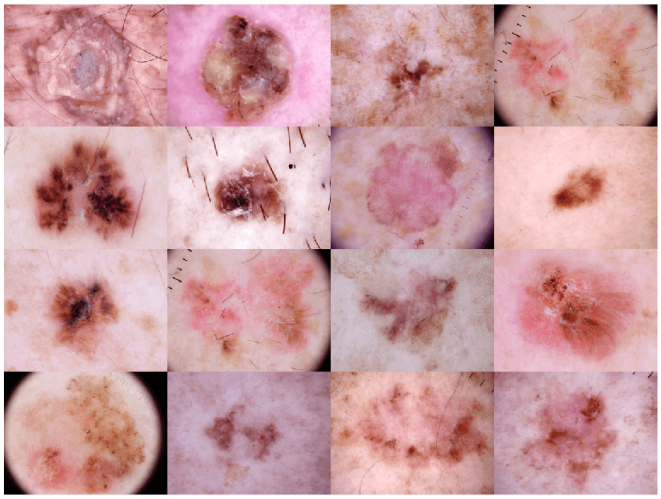


Figure 1.3: Examples of Actinic Keratosis and Intraepithelial Carcinoma (AKIEC).

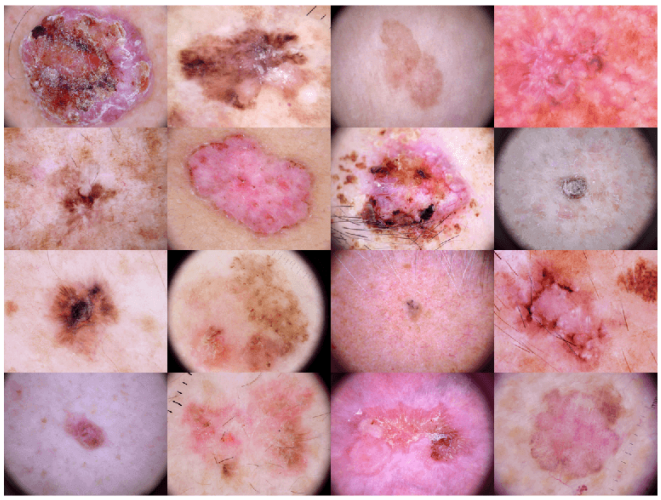


Figure 1.4: Examples of Basal Cell Carcinoma (BCC).

- **Dermatofibroma (DF):** also known as fibrous histiocytomas, they are common benign nodules caused by a proliferation of fibroblasts. Dermatofibromas can be found anywhere on the body, but most frequently arise on the legs and arms. They require no treatment and can remain unchanged for years and possibly resolve spontaneously [10].

- **Melanocytic Nevi (NV):** A melanocytic naevus or mole, is a common

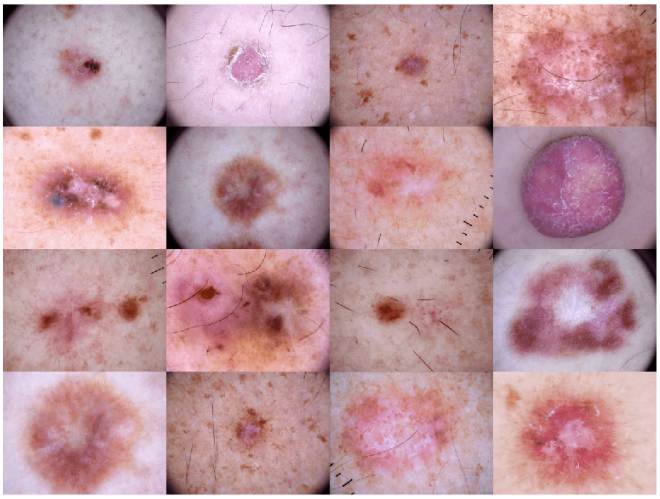


Figure 1.5: Examples of Dermatofibroma (DF).

benign skin lesion due to a local proliferation of melanocytes. They can be present at birth (a congenital melanocytic naevus) or appear later (an acquired naevus) [11].

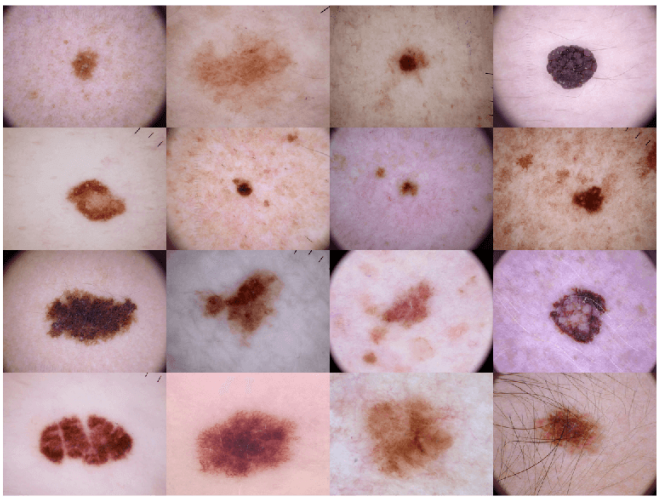


Figure 1.6: Examples of Melanocytic Nevi (NV).

- **Seborrheic Keratoses (SK):** is a non-cancerous skin tumour that appear as round warty plaques whose color ranges from light tan to black. They usually found on the chest and back, as well as on the head as people age.

They may be confused with malignant melanoma or warts. The cause of seborrheic keratoses is still unknown although they do not lead to skin cancer [12].

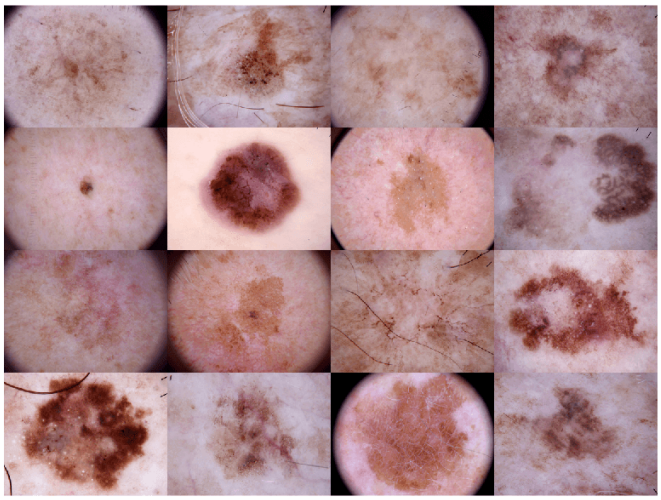


Figure 1.7: Examples of Seborrheic Keratoses (SK).

- **Vascular Lesion (VASC):** vascular naevi may be proliferative haemangiomas, which usually resolve, or vascular malformations including capillary, venous and lymphangatic types.

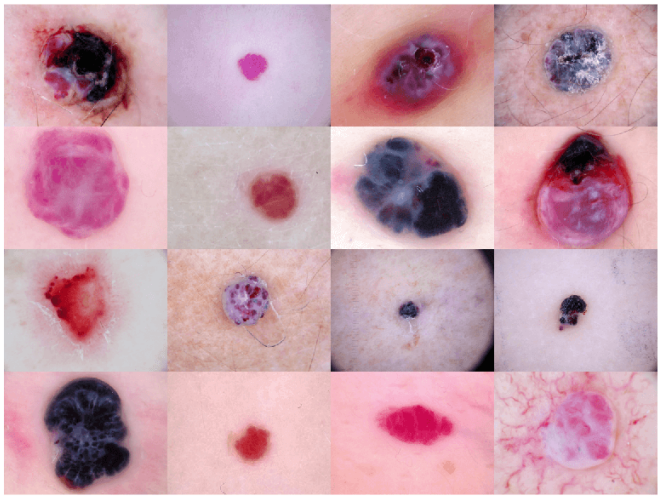


Figure 1.8: Examples of Vascular Lesion (VASC).

1.2.3 Melanoma

Melanoma is a type of skin cancer that is characterized by the uncontrolled proliferation of melanocytic stem cells that have undergone genetic transformation. Melanocytes are pigment cells that are located in the basal layer of the epidermis and produce a protein called *melanin* to protect skin cells from ultraviolet radiation [13].

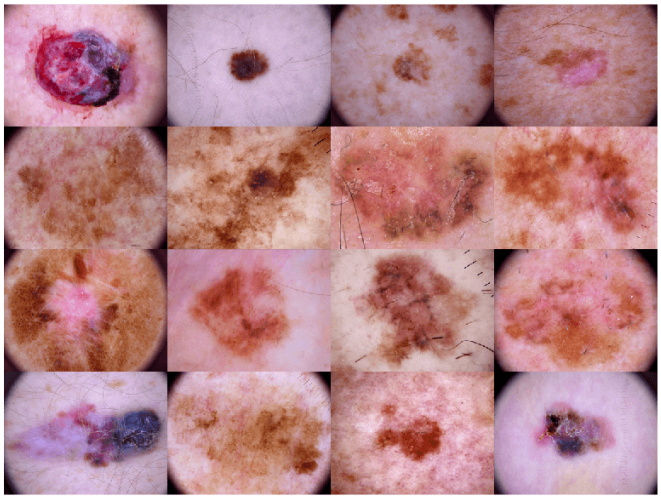


Figure 1.9: Examples of Melanoma (ML).

While melanoma is relatively uncommon in children, it is more frequently diagnosed in adults, with the average age of diagnosis having decreased in recent decades. The incidence of melanoma has been increasing globally, with a particular prevalence observed in the Caucasian population, as well as numerous Western societies [14]. This trend has affected both males and females, as depicted in Figure 1.10. In Italy, data from the Italian Association of Cancer Registries [15] estimate that there are approximately 7300 new cases of melanoma each year among men and 6700 cases among women, with the incidence having doubled in the past decade, as shown in Figure 1.11.
Melanoma is believed to be a multi-factorial disease that results from the interaction between genetic susceptibility and environmental exposure. The primary risk factors for developing melanoma include:

- UV rays exposure
- Number of melanocytic naevi or atypical naevi
- A strong family history
- Genetic susceptibility

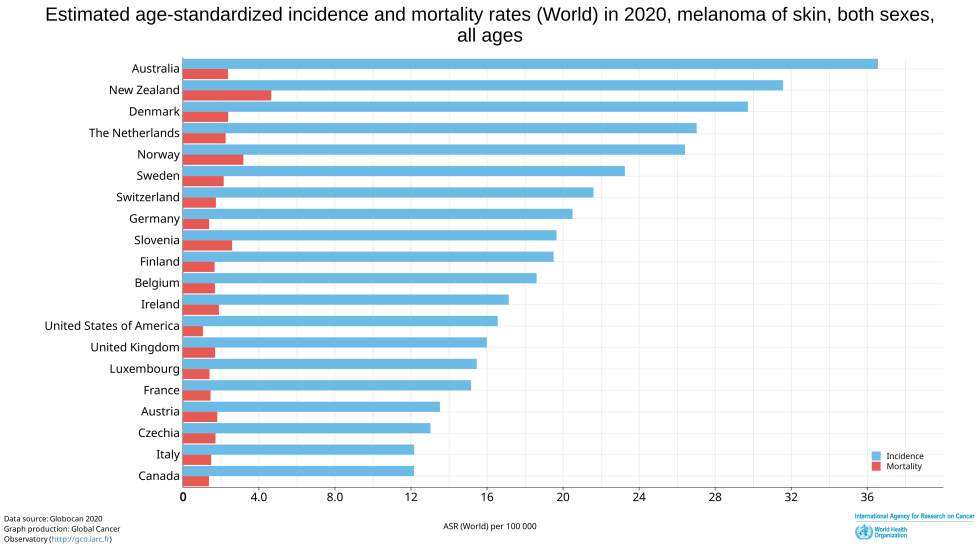


Figure 1.10: Estimated age-standardized incidence and mortality rates (World) in 2020, melanoma of skin, both sexes, all ages [5].

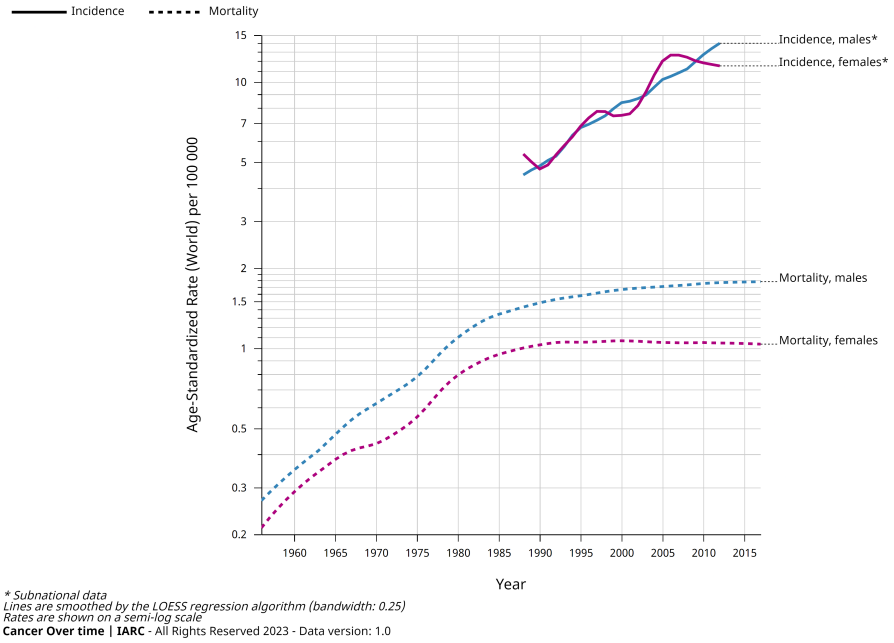


Figure 1.11: Age-standardized rate (World) per 100 000, incidence and mortality, males and females, Melanoma of skin, Italy [5].

- Previous Basal Cell Carcinoma (BCC) or Squamous Cell Carcinoma (SCC)
- Previous invasive melanoma

Non-cancerous growth of melanocytes results in moles and freckles, while cancerous growth of melanocytes typically begins as melanoma in the epidermis in what is known as the horizontal growth phase. In contrast, nodular melanoma has a vertical growth phase that is more dangerous. Once the melanoma cells have reached the dermis, they may spread to other tissues through the lymphatic system or via the bloodstream to other organs, such as the lungs or brain. This is referred to as metastatic disease or secondary spread. Although melanoma is most commonly found on the skin, it can also rarely appear on mucous membranes, such as the lips or genitals. The first sign of melanoma is often an unusual freckle or mole. During the horizontal phase of growth, a melanoma is typically flat, while it becomes thickened and raised as the vertical phase develops. Early signs in a nevus that may suggest a malignant change include discoloration, itching or ulceration, and bleeding [13].
The primary signs of melanoma can be remembered using the mnemonic ABCDEFG, which is an acronym designed to help patients and clinicians identify features in a skin lesion that may suggest an early or in situ melanoma. The acronym and its corresponding explanations are listed in Table 1.1.

Superficial Melanoma	**Nodular Melanoma**
A: asymmetry of shape	**E:** elevated
B: border irregularity	**F:** firm to touch
C: color variation	**G:** growing
D: diameter	
E: evolving (enlarging, changing)	

Table 1.1: ABCFDEG signs explanation.

While the ABCDEFG criteria have been shown to be effective in identifying potential melanomas, they are not always reliable in identifying all melanomas, particularly less common subtypes such as desmoplastic melanoma and melanoma in children [16]. Dermoscopy, a technique used by trained specialists, is more helpful in identifying malignant lesions than visual examination alone, but a skin biopsy is often required to confirm the diagnosis and determine the severity of the melanoma. Melanomas are classified based on their appearance and behavior, with those that start as flat patches and have a horizontal growth phase including:

- Superficial spreading melanoma

- Lentigo maligna melanoma and lentiginous melanoma (in sun-damaged sites)
- Acral lentiginous melanoma (on soles of feet, palms of hands or nails)

These forms of melanoma tend to grow slowly, but may progress to a vertical growth phase at any time, including:

- Nodular melanoma
- Spitzoid melanoma
- Mucosal melanoma
- Neurotropic and desmoplastic melanoma
- Spindle cell melanoma
- Ocular melanoma

Treatment for melanoma depends on various factors, such as the stage of the tumor, the severity of the condition, and the presence of other pathologies [17]. The most common intervention is surgical excision, which involves removing both the malignant and healthy tissue around the tumor. In severe cases, such as metastasis, chemotherapy combined with radiation therapy may be used to slow the spread of the tumor. Early diagnosis is crucial for successful treatment.

1.3 Dermoscopy

Dermoscopy, also known as dermatoscopy or epiluminescence microscopy, is a non-invasive diagnostic technique that has been widely adopted by dermatologists for the evaluation of suspicious skin lesions. This technique allows for the identification of lesions and differentiation between suspicious melanocytic lesions, dysplastic lesions, and melanomas, as well as the diagnosis of keratinocyte skin cancers such as basal cell carcinoma and squamous cell carcinoma. Furthermore, in recent years, dermoscopy has been utilized to diagnose a variety of dermatological conditions and disorders involving the skin's appendages, including hair and nails [18].

1.3.1 Dermatoscope and Technique

Dermoscopy is a practical and straightforward technique that allows for the visualization of subtle structural features within the reticular dermis and the recording of images for future comparison [18]. It requires minimal additional resources beyond the dermoscope itself, which consists of:

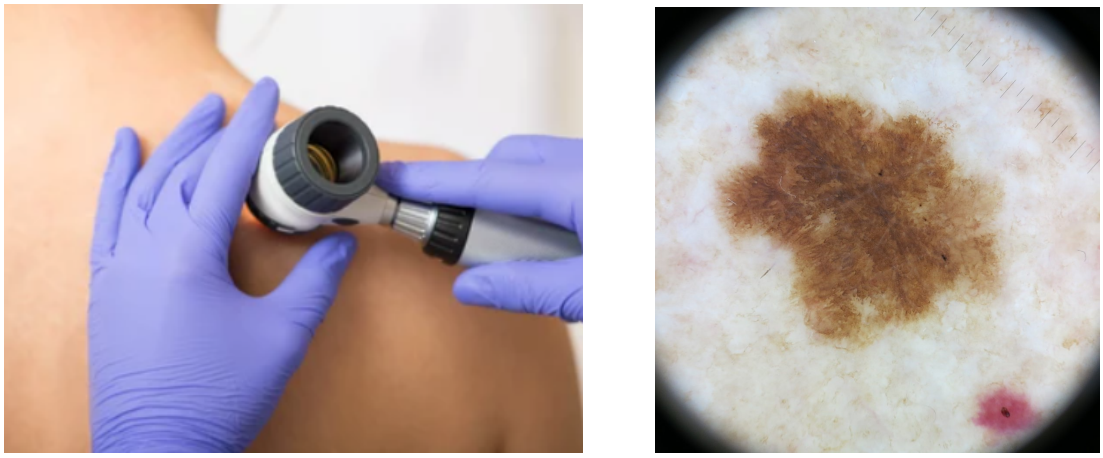


Figure 1.12: Employment of a dermatoscope (e.g., Heine Delta 20T [19]) to acquire high-resolution visual representations of skin lesions. (right). Digital dermoscopic image, which has been acquired through the utilization of a dermatoscope (left).

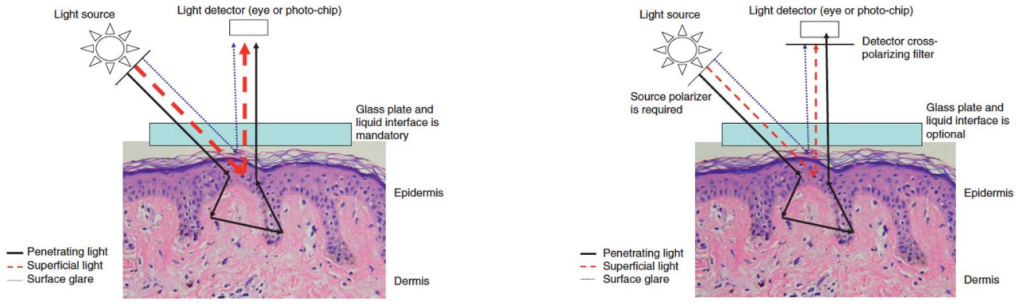
- a set of achromatic high quality lenses for 10 to 100-times magnification or even higher
- an inbuilt illuminating system
- a source of power supply (e.g. battery-powered)

When using non-polarized light [2] sources such as the sun or lamps, light is absorbed, refracted, and reflected as it hits the surface of the skin, which can create polarization and glare that reduces the visibility of underlying structures. To mitigate this issue, dermoscopy can be performed using three main techniques:

- **Non-polarised contact dermoscopy:** it uses a glass plate as a medium with a smaller refractive index to skin to minimise glare. A linkage fluid (water, oil or ECG gel which is the most commonly used nowadays) is always placed at the interface between the epidermis and the device's glass plate. It is commonly used to view superficial layers. In some situations a cling film or adhesive tape can be placed over the lesional skin to avoid inter-patient infections [20].
- **Polarised dermoscopy:** it uses polarised lenses dotate of crossed-polarizers which filter out scattered light and hence allows only light in a single plane to

[2]Non-polarised light: electromagnetic wave that vibrates in a single plane. It differs from Polarised light which vibrates in a single plane instead.

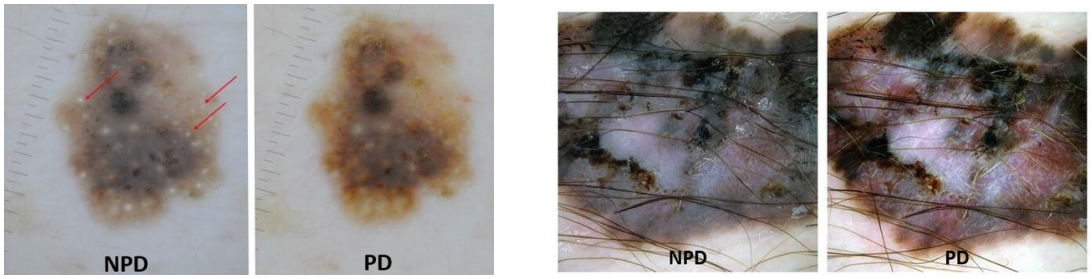
pass through it without requiring skin contact or fluid immersion (Polarised non-contact dermoscopy). It can be used to view deeper layers. However, many workers still use polarized light dermatoscopes in contact method (Polarised contact dermoscopy) with fluid systems, as it increases the penetration of light and facilitate imaging of deeper structures [20].



(a) Non-polarised optics . **(b)** Polatised optics .

Figure 1.13: Differences between non-polarised and polarised optics principles [21].

Polarised and Non-polarised methods are not mutually exclusive as they provide complementary information as shown in Figure 1.14. Moreover, when used together it is more likely to increases diagnostic accuracy and clinician confidence. As



(a) Milia like cysts on non-polarised (left) and polarised dermoscopy (right) . **(b)** Blue white veil under non-polatised (left) and polarized dermoscopy (right) .

Figure 1.14: Differences between non-polarised and polarised dermoscopic images [22].

technology has evolved, dermatoscopes have progressed from simple hand-held devices to image-capture devices with built-in digital cameras and even USB video-dermatoscopes with high-resolution cameras for real-time visualization on a computer screen. Additionally, some dermatoscopes can be attached to smart-phones for easier image capture and documentation, and some even have analytical

capabilities such as digital image analysis and incorporated artificial intelligence systems to provide automated diagnostic aid.

1.3.2 Criteria and Evaluation Models

Color and pattern structure are both important factors to consider when interpreting a potentially pigmented lesion using dermoscopy. The color of a lesion is determined by the location of melanin within the different layers of the skin, while the structure is influenced by the spatial distribution and concentration of melanin, such as whether it is present in isolation or in clusters. There are several analysis models that can be used to determine whether a lesion is benign or potentially suspect for melanoma, including semi-quantitative models like the ABCD rules, the rule of three points, and the rule of seven points, as well as qualitative diagnostic models like Menzies' method and pattern analysis. These models can be used in conjunction with other diagnostic techniques to help determine the most likely diagnosis [23].

- **ABCD method**: morphological characteristics are taken into consideration and for each one of these a value is assigned as explained in table 1.2. Finally,

Characteristics	**ABCD Method**	
	Condition	**Score**
Symmetry: lesion is divided with 2 perpendicular axes	asymmetry in both axes	2.6
	asymmetry in one axes	1.3
	total symmetry	0
Border	each abrubt border	0.1
Color: red, black, dark brown light brown, gray-bluish, white	each color found	0.5
Structure: dot, clustered globules, pigmented net, amorphous area	each structure found	0.5

Table 1.2: ABCD method: score assignment rules.

  the Total Dermoscopic Score (TDS) is calculated as sum of each individual score and a dermoscopic result is assigned as shown in table 1.3.

- **Rule of seven points:** semi-quantitative model for diagnosing melanoma using dermoscopy. It involves the assessment of seven dermoscopic characteristics, with each characteristic being divided into major and minor categories. If the total score of the characteristics is equal to or greater than three, the lesion is diagnosed as a melanoma, otherwise it is considered to be benign [23].

Total Dermoscopic Score	Diagnostic Result
score < 4.75	Benign
$4.75 <=$ **score** $<= 5.45$	Suspicious lesion
score > 5.75	Melanoma

Table 1.3: ABCD method: diagnostic result.

- **Menzies' method:** qualitative diagnostic model that involves the assessment of eleven dermoscopic characteristics that are divided into negative and positive categories. Negative characteristics, such as symmetry or a single color lesion, define the lesion as benign, while positive characteristics, such as an asymmetrical lesion, suggest the possibility of melanoma. The presence of a positive feature in conjunction with the absence of negative characteristics is sufficient for a diagnosis of cutaneous melanoma using this method [23].
- **Rule of three points:** a semi-quantitative model designed specifically for novice dermoscopists to help them improve their skills and avoid misdiagnosing melanoma. It involves the assessment of three criteria:
  - asymmetry of color and structure in one or two perpendicular axes
  - atypical network (pigmented network with irregular holes and thick lines)
  - blue–white structures

  If two or three of these criteria are present, the lesion is considered suspicious and must be excised.
- **Pattern analysis:** a method for diagnosing melanoma using dermoscopy that involves the initial analysis of the general global pattern of the pigmented lesion, followed by the assessment of local features. This method is preferred by many expert dermoscopists [23].

1.4 Computer Vision

Computer vision is an interdisciplinary scientific field that involves the development of systems that can understand and interpret digital images or videos in order to perform tasks that the human visual system is capable of. This includes acquiring, processing, analyzing, and understanding digital images to extract high-dimensional data from the real world, resulting in numerical or symbolic information, such as decisions. Subfields of computer vision include object detection, object recognition, image classification, image segmentation, image restoration, and

etc. Image processing, which encompasses various techniques and algorithms for manipulating digital images for tasks such as pattern recognition, feature extraction, and image enhancement, can be considered a subfield of computer vision. It has been widely applied in commercial and industrial fields, medical applications, and scientific research and continues to be a growing area of study [24].

1.4.1 Digital Image

A digital image is a two-dimensional representation of a visual scene, typically consisting of a finite set of pixels arranged in rows and columns [25]. Each pixel, or picture element, is the smallest unit of an image and is characterized by its coordinate position within the image matrix, with the x-axis increasing from left to right and the y-axis increasing from top to bottom. The main components of a digital image include:

- **Size:** it refers to the number of pixels in the image, commonly in terms of height and width.
- **Intensity:** it is the brightness of each pixel.
- **Color space:** it determines how colors are represented in the image and can be expressed using various color models such as grayscale, RGB, or CMYK.
- **Channel:** they are attributes of the color space and can vary in number depending on the color model used. For example, grayscale images have a single channel, while RGB images have three color channels: red, green, and blue.

The dynamic range of an image, which describes the range of intensity values that can be represented, is determined by the number of bits used to represent the intensity of each pixel and can be expressed mathematically by 2^{N-1} where N represents the number of bits. A standard digital photo uses an 8-bit range, which allows for 256 possible intensity values ranging from 0 (black) to 255 (white). Grayscale images use a single 8-bit intensity range, while RGB images use 8-bit intensity ranges for each of their three color channels.

1.5 Artificial Intelligence (AI)

Artificial intelligence (AI) is the umbrella term for techniques that enable machines, such as computers, robots, and autonomous vehicles, to perform tasks that are characteristic of human cognitive abilities, such as learning and problem-solving. AI is used in a wide range of applications, including search engines (e.g., Google),

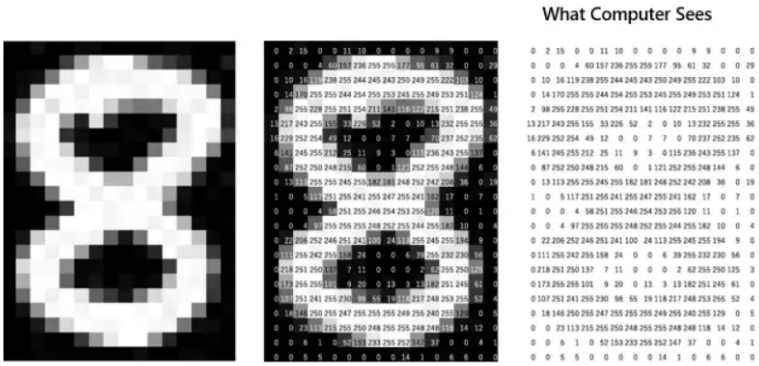


Figure 1.15: Image representing the number eight seen by Human Vision (left) respect to Computer Vision (right) [26].

recommendation systems (used by YouTube, Amazon, and Netflix), speech recognition (e.g., Siri and Alexa), self-driving cars (e.g., Tesla), and many others. With the increasing availability of data, storage capacity, and computational power, AI has become increasingly prevalent in various fields, including industry and academia, and has even outperformed human beings in certain highly mathematical and statistical problems. AI encompasses a range of methods, including genetic algorithms, fuzzy models, and machine learning [27].

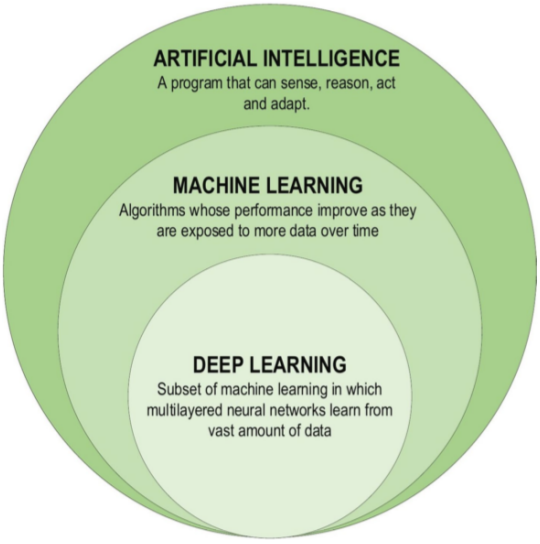


Figure 1.16: Relationship between Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) [28].

1.5.1 Machine Learning and Deep Learning

Machine learning (ML) is a subset of artificial intelligence (AI) that refers to the ability of intelligent systems to learn and extract meaningful patterns and relationships from problem-specific training data, which consists of examples and observations. The ultimate goal of ML is to automate the process of solving associated tasks and make reliable and repeatable decisions. ML models have been effective in tasks involving high-dimensional data, such as classification, regression, and clustering, to create analytical models that generate predictions, rules, recommendations, or similar outcomes. As a result, ML algorithms have been applied in a variety of areas, including fraud detection, credit scoring, speech and image recognition, and natural language processing (NLP). In recent decades, the field of ML has seen significant advances with the evolution of artificial neural networks (ANNs) towards deeper neural network architectures with improved learning capabilities, known as deep learning (DL). In contrast, simple ANNs (e.g., feedforward neural networks, shallow autoencoders) and other ML algorithms (e.g., support vector machines, decision trees) can be classified as shallow machine learning [27]. The hierarchical relationship between AI, ML, and DL is illustrated in Figure 1.16. DL neural networks are based on a computational model inspired by the functioning of the biological brain. Like biological neurons in the brain, a large number of interconnected nodes, called artificial neurons, form multiple hidden layers, organized in deeply nested network architectures as shown in Figure 1.17. This allows DL neural networks to use advanced operations (e.g., convolutions) or

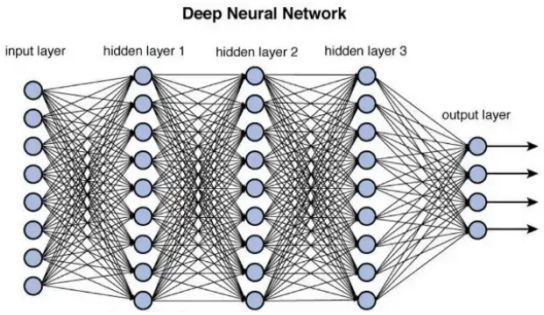


Figure 1.17: Deep Neural Network architecture with multiple layers [29].

multiple activations in one neuron, rather than using a simple activation function. These characteristics enable DL neural networks to be fed raw input data and automatically discover a representation needed for the corresponding learning task. The learned features can then be used to infer the nature of an unseen input. In contrast, shallow ML heavily relies on carefully defined features, and therefore its performance is dependent on a successful feature extraction process, as shown in Figure 1.18. This process can be time-consuming, labor-intensive, and inflexible,

as it often requires a lot of domain expertise and application-specific engineering. DL neural networks overcome this limitation of handcrafted feature engineering by having the capability of automated feature learning, which allows them to extract discriminative feature representations with minimal human effort [27]. DL is particularly effective in dealing with large, noisy, and unstructured data, and is therefore useful in domains with large and high-dimensional data. This is why DL neural networks tend to outperform shallow ML models and traditional data analysis approaches for most applications involving text, image, video, speech, and audio data. However, for low-dimensional data inputs, especially in cases of limited training data availability, shallow ML can still produce superior results, which may even be more interpretable than those generated by DL neural networks. Based on

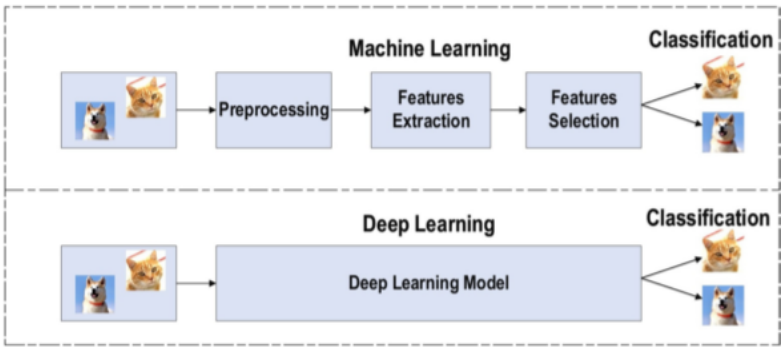


Figure 1.18: Difference between Machine Learning and Deep Learning [28].

the characteristics of the problem and the available data, machine learning (ML) and deep learning (DL) techniques can be classified into three major categories:

- **Supervised learning:** this technique requires a training dataset that includes input examples and labeled output or target values. The input-output pairs in the training set are used to calibrate the parameters of the ML model. Once the model has been trained, it can be used to predict the target variable given new or unseen input data points. Supervised learning can be further divided into regression problems, where a numeric value is predicted (e.g., number of users), and classification problems, where the prediction result is a categorical class [27].

- **Unsupervised learning:** in this case, the system is expected to detect patterns within the dataset without the need for labels. Therefore, the training data consists only of input variables and the goal is to find structural information of interest, such as groups of elements that share common properties (known as clustering) or data representations that are projected from a high-dimensional space into a lower one (known as dimensionality reduction).

- **Reinforcement learning:** instead of providing input-output pairs, the network is given the current state of the system, the ultimate goal, and a list of allowable actions and environmental constraints for the outcomes. The goal is achieved through trial and error, using a reward system to maximize the reward.

1.5.2 AI Applications in Dermatology

Suspicious skin lesions can be challenging to accurately evaluate and identify, particularly due to the variable appearance of melanoma, which can lead to misdiagnosis or unnecessary surgical excision. In recent years, the use of artificial intelligence (AI) in the field of dermatology has been growing, driven by the availability of more data and images of skin lesions and technological advancements. While AI is not expected to replace medical experts in the near future, it can offer a second opinion and be used to screen out benign lesions, such as symmetrical melanocytic nevi. In the future, as databases continue to expand with more patients and images with lesion-specific labels, more complex algorithms will be developed, leading to improved accuracy in the detection of potentially malignant skin lesions.
In the past, the adoption of sophisticated specialized systems for image acquisition and storage has been slow due to cost and inconvenience, preventing the latest generation of dermatoscopes from being accessible to all patients worldwide. However, with the rise of the mobile revolution and digitalization in healthcare, smartphone-based digital cameras and dermatoscopes are entering the professional market, increasing accessibility to healthcare in patient treatment and screening. During the COVID-19 pandemic, the use of digital imaging in *Teledermatology* [3] applications has steadily increased. The possibility of direct teleconsultation for concerning lesions between patients and dermatologists has the potential to improve accessibility to care. However, patient-acquired images often have low quality due to poor lighting conditions, blurriness, and out-of-focus effects. Improving and enhancing the quality of such poor images is important as high-quality, detailed images are essential for accurate diagnosis of skin lesions [30].

1.6 Convolutional Neural Network (CNN)

Convolutional neural networks (CNNs) are a type of deep neural network that is particularly popular for tasks related to computer vision and speech recognition. One of the main advantages of CNNs is that they can automatically detect significant

[3]Teledermatology: sub-field of Telemedicine. It refers to various aspects of dermatology delivered using modern information and communications technology (ICT)

features without any human supervision. CNNs are inspired by the structure and function of the human visual cortex, consisting of alternating layers of simple and complex cells. While there have been several proposed changes to the CNN architecture in recent years with the aim of improving image classification accuracy or reducing computation costs, the most popular base architecture consists of several modules, each made up of a group of convolutional and pooling layers, stacked on top of each other. These representations are then fed into one or more fully connected layers, and the final fully connected layer outputs the class label [28]. For example, in object recognition tasks using images, the first few layers of the CNN are responsible for extracting basic features from the input images, such as edges and curves. These basic features are then gradually aggregated into more complex abstract features in the final layers, resembling the objects of interest. The generated features are then used for prediction purposes, allowing the CNN to recognize objects of interest in new images or differentiate between different image classes.

1.6.1 Network Architecture

The CNN architecture consists of various types of layers, each with its own unique characteristics and functions:

- **Convolutional Layer:** is considered a crucial component of CNN models. This layer is composed of convolutional filters, also known as kernels, which are responsible for learning to extract important features from the input image. These kernels are convolved with the input image to produce the output feature map.

- **Pooling Layer:** it takes the large-size feature map generated by previous convolutional operations and reduces its size while preserving the majority of relevant information.

- **Activation Function:** also referred to as the transfer function, plays a crucial role in mapping the input to the output by defining how the weighted sum of the input, along with its bias (if present), is transformed into an output by the nodes in a layer of the network. Non-linear activation layers are typically used after all layers with weights (such as FC layers and convolutional layers) in the CNN architecture. The non-linear behavior of the activation layers allows the model to learn more complex relationships between the input and output.

- **Fully Connected (FC) Layer:** is composed of neurons that are connected to all neurons in the previous layer, following the conventional structure of a multi-layer perceptron (MLP).

- **Loss Function:** is used in the output layer to calculate the predicted error that occurs during the training phase. This error represents the difference between the actual output and the predicted output of the model. The loss function is then optimized during the learning process of the CNN to improve the performance of the model.

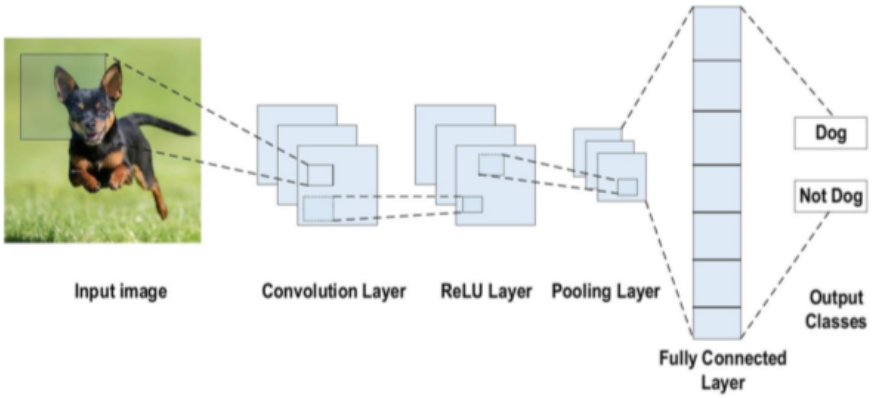


Figure 1.19: Basic architecture and general functioning of a Convolutional Neural Network (CNN) [28].

1.6.2 Backpropagation

One key component of CNNs is the ability to train them using backpropagation, a method for efficiently computing the gradients of the loss function with respect to the model parameters. Backpropagation involves forward propagating the input through the network to generate an output, and then propagating the error back through the network to update the model parameters. This is done through the use of the chain rule, which allows the gradients to be computed recursively from the output layer down to the input layer. Overall, backpropagation is a crucial component of CNNs, allowing them to be trained effectively on large datasets and achieving state-of-the-art performance on a variety of tasks [28].
As an example, consider the task of image classification using a CNN. The input to the network is an image, and the output is a probability distribution over a set of classes. The goal of training is to adjust the model parameters such that the output of the network is as close as possible to the true label of the image. During training, a batch of images is fed through the network, and the output is compared to the true labels using a loss function such as cross-entropy loss. The gradient of the loss with respect to the model parameters is then computed using backpropagation and the parameters are updated using an optimization algorithm

such as Stochastic Gradient Descent (SDG) or Adam. This process is repeated for multiple epochs until the model converges.

1.6.3 Hyperparameters

The learning process for convolutional neural networks (CNNs) involves the selection of an appropriate learning algorithm, also known as an optimizer, and the use of various enhancements to improve the output of the model. Loss functions, which aim to minimize the error between the predicted and actual output, are central to all supervised learning algorithms. Gradient-based learning algorithms are commonly used in CNNs, and involve repeatedly updating the network parameters through each training epoch in order to minimize the training error. This process is performed through back-propagation, in which the gradient at each neuron is propagated back to all neurons in the preceding layer. There are several variations of gradient-based learning algorithms, including batch gradient descent, stochastic gradient descent, and mini-batch gradient descent. In addition to these algorithms, various enhancement techniques can be used to further improve the training process, such as momentum, Adagrad. It is important to carefully select the learning rate, which determines the step size of the parameter updates, in order to avoid negatively impacting the learning process [28]. However, gradient-based learning algorithms can be susceptible to getting stuck in local minima rather than finding the global minimum solution, particularly if the solution space is not convex. To address this issue, techniques such as simulated annealing and early stopping can be employed.

1.6.4 Regularization

Overfitting is a common issue in convolutional neural network (CNN) models, where the model performs exceptionally well on training data but poorly on unseen test data. Underfitting occurs when the model does not learn enough from the training data, while a just-fitted model performs well on both training and testing data. There are several techniques that can be used to address overfitting, such as dropout, which involves randomly dropping neurons during training, and drop-weights, which involves randomly dropping connections between neurons. Batch normalization can also help to reduce the risk of overfitting, as it helps to ensure the performance of output activations and reduce the internal covariate shift of activation layers. Other advantages of batch normalization include preventing the vanishing gradient problem, improving weight initialization, reducing the time required for network convergence, and decreasing the dependence on hyperparameters during training [28].

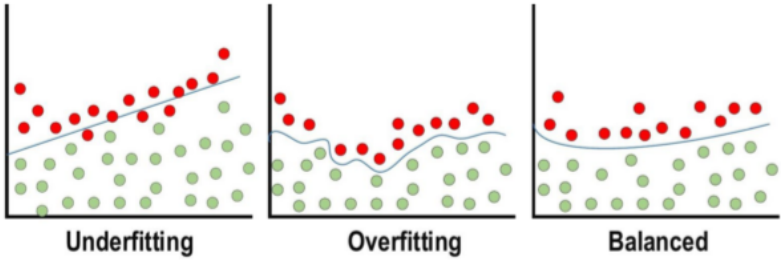


Figure 1.20: Graphic representation of a well-balanced model respect to Overfitting and Underfitting phenomena [28].

1.6.5 Improvements

Based on experiments with various deep learning (DL) applications, several approaches may be effective in improving the performance of convolutional neural networks (CNNs). One such approach involves expanding the dataset through techniques such as data augmentation, or using the transfer learning, which is a technique that enables models to be trained on large-scale image datasets and then specialized for specific tasks using problem-specific datasets. Additionally, increasing the duration of training may also contribute to enhanced performance. Another approach that may be beneficial is increasing the depth or width of the CNN model, as well as implementing regularization techniques and fine-tuning hyperparameters [28].

1.7 Generative Adversarial Network (GAN)

Generative adversarial networks (GANs) are a type of machine learning algorithm that has the potential to revolutionize domains where new content or product configurations are constantly being created, such as art, music, and fashion design, or where content is transformed from one representation to another, such as text to image for product descriptions. However, GANs also pose significant risks with societal implications if they are abused for malicious purposes. In particular, the generation of 'deepfake' content, such as abusive speeches and misleading news to manipulate public opinion or distort financial markets, is a concern [31].

1.7.1 Network Architecture

Generative adversarial networks (GANs) are a class of generative models that aim to learn the probability distribution of a dataset such that the model can generate new, randomly varied samples from the distribution. GANs consist of two sub-networks: a Generator and a Discriminator. The Generator network, denoted

as G, captures the distribution of an input random noise vector z and generates new examples $G(z)$ that closely resemble samples from the target domain. On the other hand, the Discriminator network, denoted as D, attempts to differentiate between real samples from the target domain and artificially generated samples produced by the Generator. These two networks are trained simultaneously through an objective function based on a min-max game optimization problem, in which the Generator tries to minimize the loss function while the Discriminator tries to maximize it. The parameters of the networks are updated through backpropagation until the Discriminator is unable to distinguish between real and fake samples. The ultimate goal of the Generator is to produce realistic-looking samples, while the Discriminator strives to become increasingly proficient at detecting fake samples [31]. The objective function of a GAN can be expressed as:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[1 - \log D(G(z))]$$

where E_x is the expected value over all real data samples, $D(x)$ is the probability estimate of the Discriminator that the sample x is real, $G(z)$ is the output of the Generator for a given random noise vector z as input, $D(G(z))$ is the Discriminator's probability estimate that the fake generated sample is real, and E_z is the expected value over all random inputs to the Generator.

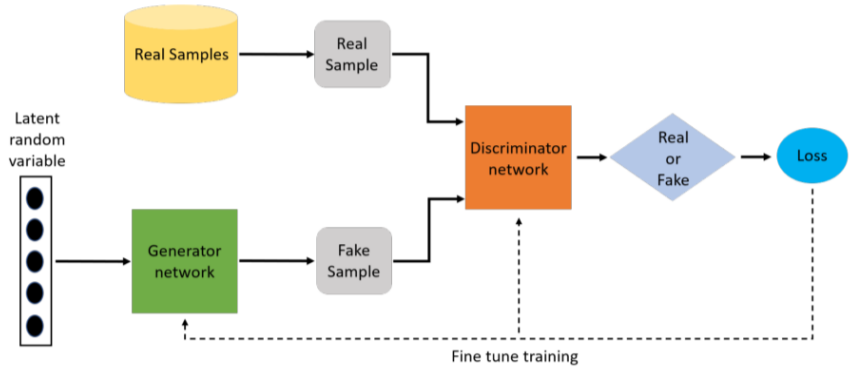


Figure 1.21: Basic architecture and general functioning of a Generative Adversarial Network (GAN) [31].

1.8 Super-Resolution Image Reconstruction

The applications of super-resolution techniques have been broad and encompass several areas such as video processing, satellite imaging, and medical imaging. The recording of digital images often results in a loss of spatial resolution due to a range

of factors, including optical distortions, motion blur, sensor or transmission noise, and insufficient sensor density. As a result, the recorded images frequently exhibit blurring, noise, and aliasing effects. Super-resolution image reconstruction, also referred to as resolution enhancement, is a subset of techniques and algorithms used in computer vision and image processing for the purpose of restoring HR images from LR images. In the field of medical imaging, obtaining high-quality images is crucial for enabling physicians to make accurate diagnoses, as high-resolution (HR) images provide more features and details than low-resolution (LR) images. In recent years, there has been a marked increase in interest in SR methods, particularly in the domains of computed tomography (CT) and magnetic resonance imaging (MRI). The ability to acquire multiple images while still facing limitations in resolution quality has motivated the exploration of SR techniques as a means of improving diagnostic accuracy by enhancing image resolution. As such, SR has become a crucial tool in the field of medical imaging, with a growing interest in its application to various imaging modalities.
Super-resolution image reconstruction comprises both image upsampling and image restoration. The objective of image upsampling is to increase the size of an image while preserving as much information as possible, while the aim of image restoration is to recover a degraded image, such as a blurred or noisy image, without changing its size. In this way, super-resolution image reconstruction can effectively address the limitations of the imaging system and enhance the performance of various digital image processing applications [32].

1.8.1 Classic Interpolation Methods

Image interpolation refers to the process of resizing a digital image and is commonly used in computer vision applications. Traditional interpolation methods, such as nearest-neighbor, bilinear, and bicubic interpolation, are easy to understand and implement, and are therefore still commonly used in super-resolution-based approaches [33]. However, these interpolation-based upsampling methods only improve image resolution based on the existing information in the image and do not introduce any new information. As a result, they may introduce side effects, such as noise amplification and blurring, into the super-resolution model.

Nearest Neighbor Interpolation

Nearest-neighbor interpolation is a simple interpolation algorithm that selects the value of the nearest pixel for each position to be interpolated without considering any other pixels within the grid. This method is often used in real-time applications due to its fast execution, but it generally produces low-quality results.

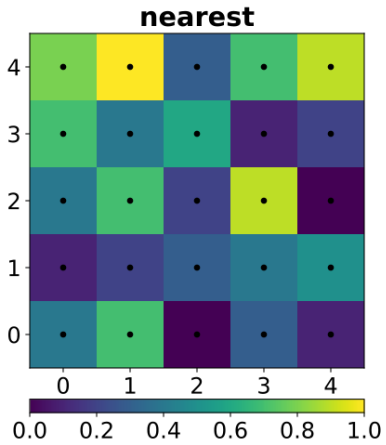


Figure 1.22: Nearest-neighbor interpolation on the square [0,4] × [0,4] consisting of 25 unit squares patched together. Colour indicates function value. The black dots are the locations of the prescribed data being interpolated [34].

Bilinear Interpolation

Bilinear interpolation is a two-step process that involves performing linear interpolation in one direction of the image and then again in the other direction. The resulting approximation is independent of the order in which the interpolation is performed. While each step is linear, the overall process results in a quadratic interpolation considering the closest 2x2 neighborhood of known pixel values surrounding the unknown pixel's computed location. Bilinear interpolation produces better results than nearest-neighbor interpolation while maintaining relatively fast speed [33].

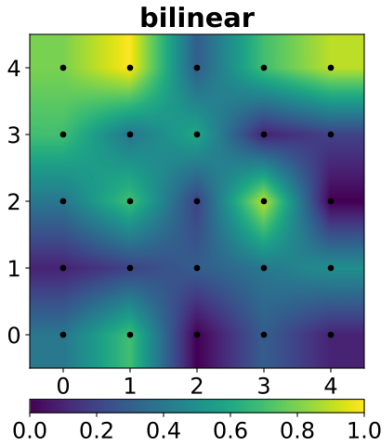


Figure 1.23: Bilinear interpolation on the same dataset as above [34].

Bicubic Interpolation

Bicubic interpolation performs cubic interpolation on each of the two dimensions of the image. It is preferred over bilinear or nearest-neighbor interpolation when speed is not a primary concern. Bicubic interpolation takes into account a 4x4 grid of pixels (16 pixels in total) and produces smoother results with fewer interpolation artifacts, but it is slower than bilinear interpolation. The bicubic interpolation with anti-aliasing is currently the most widely used method for constructing super-resolution datasets (i.e., degrading HR images to corresponding LR images) and is also commonly used in pre-upsampling super-resolution frameworks [33].

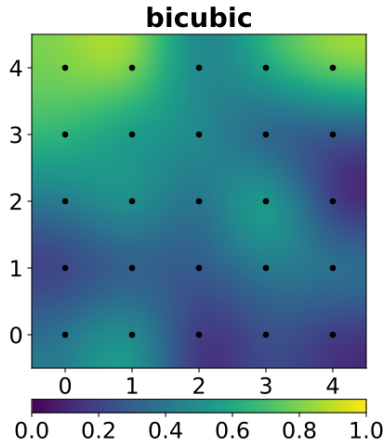


Figure 1.24: Bicubic interpolation on the same dataset as above [34].

1.8.2 AI State-of-the-Art Generative Methods

In recent years, the rapid development of deep learning techniques has led to the active exploration of deep learning-based super-resolution (SR) models, which have achieved state-of-the-art performance on various SR benchmarks. A wide range of deep learning methods have been applied to SR tasks, ranging from early convolutional neural network (CNN)-based methods (e.g., SRCNN) to more recent SR approaches using generative adversarial networks (GANs) (e.g., SRGAN). In general, the family of SR algorithms that use deep learning techniques differ from one another in terms of network architectures, loss functions, learning principles, and strategies. These differences can significantly impact the performance and efficiency of the SR model.

SRCNN

SRCNN is a deep convolutional neural network (CNN) developed for the task of single image super-resolution (SR). The SRCNN model is designed to take a low-resolution input image and produce a higher-resolution output image through the use of convolutional neural network (CNN) layers. It consists of three main components: a feature extraction layer, a non-linear mapping layer, and a reconstruction layer. The feature extraction layer uses convolutional filters to extract high-level features from the input image, while the non-linear mapping layer uses a non-linear activation function to map the extracted features to a higher-dimensional space. The reconstruction layer then uses deconvolutional filters to upsample the mapped features and produce the final high-resolution output image. SRCNN has been shown to produce high-quality SR results on a variety of datasets and has been used in numerous applications, such as medical imaging, surveillance, and remote sensing. It has also inspired the development of several other CNN-based SR models [35].

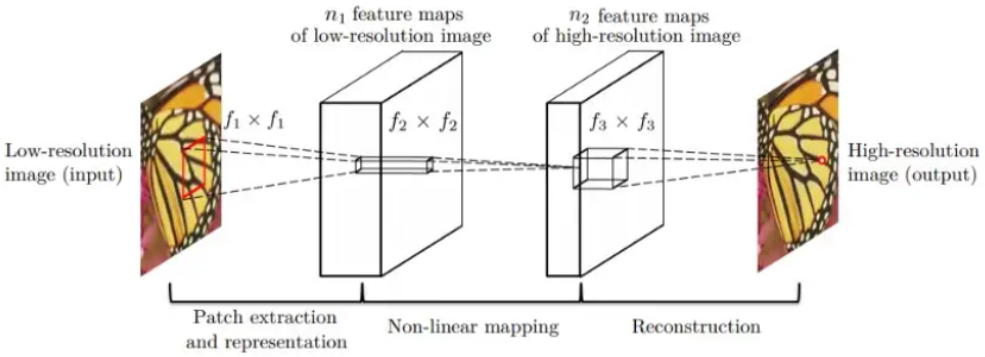


Figure 1.25: SRCNN architecture [35].

SRGAN

Following the success of SRCNN, researchers began to explore other approaches to SR that leveraged the power of deep learning. One such approach was the use of generative adversarial networks (GANs) for SR, which consist of a generator network and a discriminator network, which are trained in an adversarial manner. In the SRGAN model, the generator network is trained to generate high-resolution images from low-resolution input images, while the discriminator network is trained to distinguish between real high-resolution images and synthetic high-resolution images generated by the generator. SRGAN has been shown to produce high-quality SR results and has been used in a variety of applications [36].

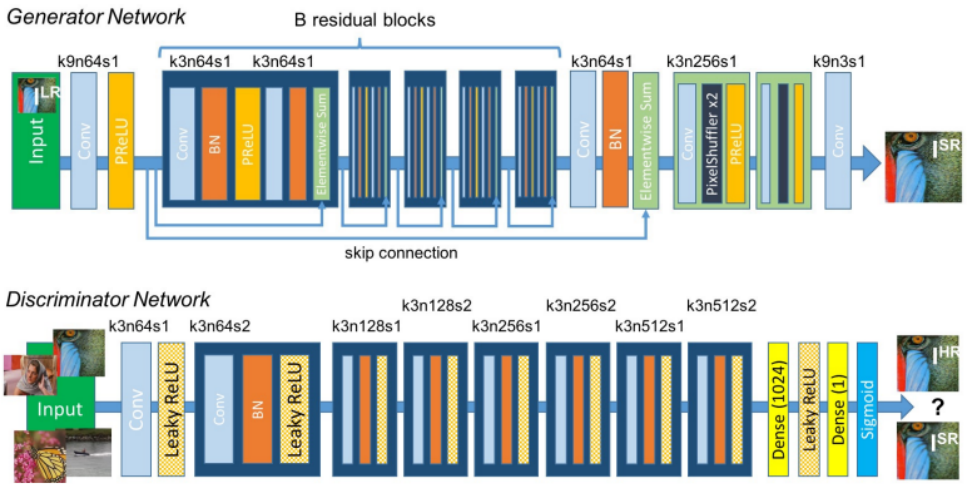


Figure 1.26: SRGAN architecture [36].

DCGAN

DCGANs utilize deep convolutional neural networks (CNNs) for both the generator and discriminator models. This represents a departure from the original GAN architecture, which utilized multi-layer perceptrons (MLPs). The use of CNNs, which are known to be effective for image processing tasks, allowed for the development of complex and high-quality generators through the enforcement of certain constraints. Three key features of the DCGAN architecture are the use of transposed convolutions in the generator, which allow for a successive increase in representation at each layer as the model maps from a low-dimensional latent vector to a high-dimensional image, the use of batch normalization in both the generator and discriminator, and the use of ReLU activation in the generator and LeakyReLU activation in the discriminator. Additionally, the DCGAN architecture utilizes the Adam optimizer rather than the traditional stochastic gradient descent with momentum. These modifications enabled stable training of the DCGAN model [31].

ProGAN

ProGAN is based on the idea of progressively increasing the resolution of the generated images during training, allowing the generator network to learn the high-frequency details of the image distribution. This results in the generation of high-resolution synthetic images. ProGAN has been used for tasks such as image generation and style transfer, and has inspired the development of other models such as StyleGAN and BigGAN [31].

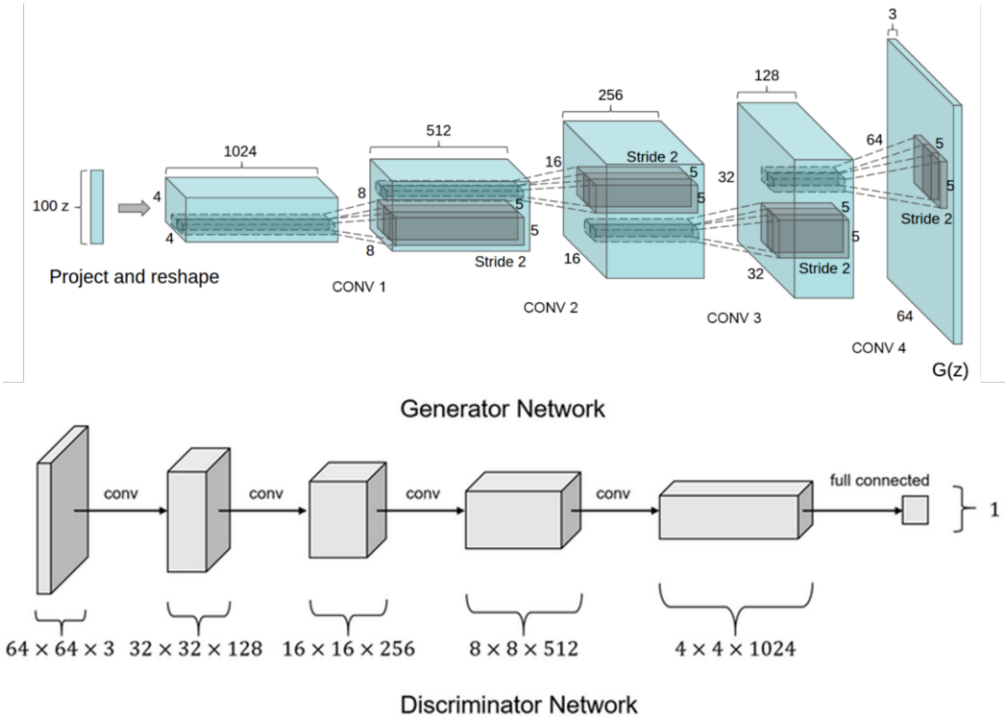


Figure 1.27: DCGAN architecture [31].

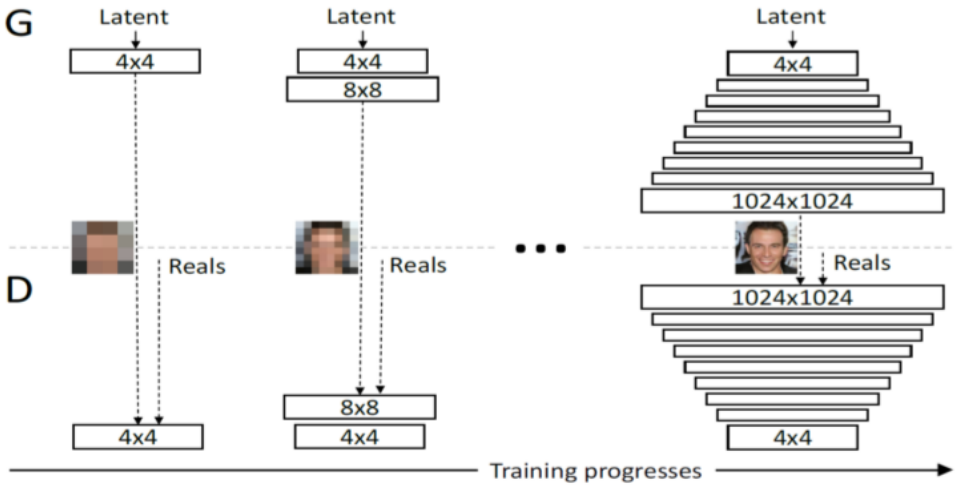


Figure 1.28: ProGAN architecture [31].

Diffusion Models

Diffusion models, also referred to as diffusion probabilistic models, are a type of latent variable model that uses Markov chains and variational inference to learn

the latent structure of a dataset. They do this by modeling the way in which data points move or 'diffuse' through the latent space. In computer vision, diffusion models are often used for tasks such as image denoising, inpainting, super-resolution, and image generation. For example, a diffusion model could be trained to denoise images that have been blurred with Gaussian noise by learning to reverse the diffusion process. Alternatively, an image generation model could start with a random noise image and, after being trained on natural images, be able to generate new natural images by reversing the diffusion process. These models have a wide range of applications, and can be useful for improving the performance of various machine learning tasks [37].

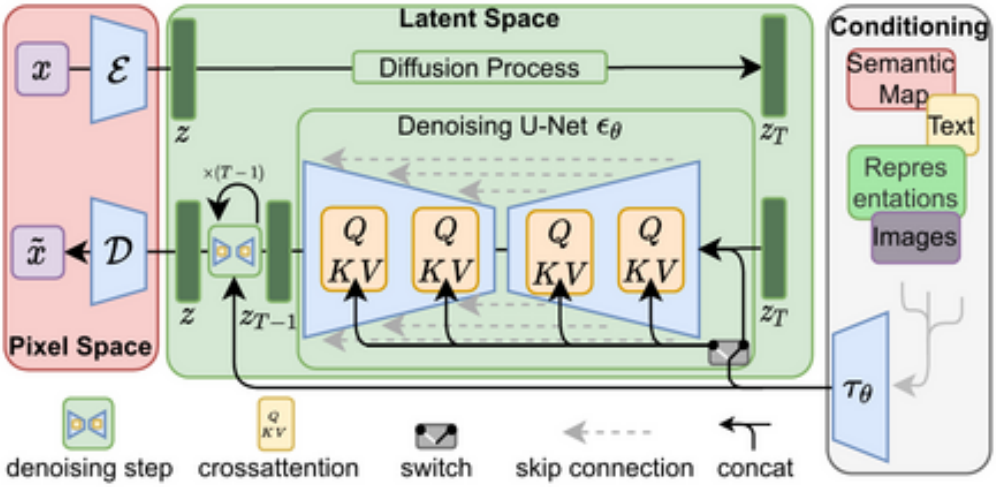


Figure 1.29: Stable Diffusion architecture [37].

Chapter 2

Materials and Methods

2.1 Dataset

In this study, several skin datasets were employed to train and evaluate Generative Adversarial Networks (GANs) models. The ISIC Archive dataset, which contains a substantial number of images partitioned into 7 classes, was predominantly employed during the training stage. On the other hand, diverse datasets were employed as Test Sets during the evaluation phase. These datasets were chosen as they present dissimilar textures, content, and characteristics in their images, providing a means of validating the model's performance when the input data differs from what was observed during the training phase.

Dataset	**# of images**	**Usage**
ISIC Archive	30604	Training and Validation Set
ISIC Test	9897	Test Set
CC ISIC Test	9897	Test Set
Atlas	2013	Test Set
Novara GQ	60	Test Set
Novara BQ	90	Test Set
Nurugo	222	Test Set
PH2	200	Test Set

Table 2.1: A comprehensive list of all the datasets that were utilized, including information on the number of samples present in each dataset and the specific usage of each dataset within the study.

2.1.1 ISIC

The International Skin Imaging Collaboration (ISIC) is a partnership between academic centers and companies from around the world that aims to reduce melanoma-related deaths and unnecessary biopsies by improving the accuracy and efficiency of melanoma early detection. To support these efforts, ISIC maintains the ISIC Archive, an open source platform that contains approximately 70000 publicly available images of skin lesions along with metadata describing additional attributes on an image level. These images were used to train GAN models in this study to reconstruct high-resolution skin lesion images from low-quality ones. The ISIC Archive is an important resource for researchers and clinicians working on skin lesion diagnosis and classification, as it provides a large and diverse dataset of images that can be used to develop and validate machine learning models and other diagnostic tools.[38].
The dataset utilized in the training stage of this study consists of approximately 30000 dermoscopic skin images, divided into 7 categories or labels:

1. **Actinic Keratosis and Intraepithelial Carcinoma (AKIEC):** 1066 images
2. **Basal Cell Carcinoma (BCC):** 3252 images
3. **Dermatofibroma (DF):** 246 images
4. **Keratosis Like (KL)** 2333 images
5. **Malignant Melanoma (ML):** 5375 images
6. **Melanocytic Nevi (NV):** 18079 images
7. **Vascular Lesions (VASC):** 253 images

Following the implementation of a thorough processing pipeline that resulted in the production of a final Training Set, along with a corresponding Validation Set, a Test Set was created using the images that were discarded from the ISIC Archive. The images in the Test Set were not used at any point during the model's training or validation, providing a robust means of assessing the generalizability of the model.

2.1.2 Color Constancy ISIC

The Color Constancy (CC) ISIC dataset consists of images identical to those in the ISIC Archive, but they undergo processing by the DermoCC-GAN algorithm [39]. This approach addresses the challenge of variation in the light source during image acquisition by formulating the color constancy task as an image-to-image

translation problem. As a result, the DermoCC-GAN algorithm is capable of generating a normalized color version of the images, thereby enhancing the accuracy of dermatological diagnoses.

2.1.3 Atlas

The Atlas dataset consists of over 2000 clinical and dermoscopy color images, along with structured metadata specifically designed for training and evaluating computer-aided diagnosis (CAD) systems. This dataset is useful for training and evaluating CAD systems due to the inclusion of both images and corresponding metadata [40].

2.1.4 Novara

The dataset in question is comprised of images obtained by the dermatologists at the Dermatology Section of Azienda Ospedaliera Universitaria Maggiore della Carità in Novara. These images were captured using the HEINE Delta 20T dermatoscope. The images are 24-bit RGB color images with a resolution of 2272x1704 pixels. The dataset is splitted into 2 different version, respectively called 'Novara Good Quality (GQ)' and 'Novara Bad Quality (BQ)' based on their quality images. The Novara GQ dataset, which comprises of images that have been meticulously selected for their exceptional resolution, is considered to be among the most superior datasets available.

2.1.5 Nurugo

This dataset was obtained using the Nurugo Derma, a smartphone lens that allows for the analysis of skin features through a 38x magnification lens. However, this tool has several limitations, including glare effect artifacts due to skin reflection and shadow effects in the corners of the images. In this study, these limitations were considered when using the dataset.

2.1.6 PH2

The dermoscopic images within the PH2 dataset were acquired at the Dermatology Service of Hospital Pedro Hispano, located in Matosinhos, Portugal, using the Tuebinger Mole Analyzer system with a magnification of 20x [41]. These images are 8-bit RGB color images with a resolution of 768x560 pixels. The database consists of 200 dermoscopic images of melanocytic lesions, including:

- **Common Nevi:** 80 images

- **Atypical nevi:** 80 images
- **Melanomas:** 40 images

2.2 Preprocessing: Data Preparation

2.2.1 Data Analysis

Dataset analysis is one of major aspect to contribute in a good results of neural network. Digital medical images commonly have complex nature due to different dimension resolutions, acquisition conditions and irregular lesion borders. Specifically, dermoscopic images of the ISIC Archive present various size resolutions (ranging from 600x450 to 6661x4499), photographic angles and lightening conditions. Moreover, most of the images may also be affected from artifacts such as black frame, hair strands, ruler measures and air bubbles, as shown in Figure 2.1. As first step of pre-processing pipeline, images from ISIC Archive were visually observed and only those heavily compromised due to artifacts or extremely poor quality (e.g., out-of-motion, blurry effects) were discarded. However the main goal is to maintain the variability of the dataset as high as possible in order to increase the network generalization capabilities in real-world scenario, hence, including images with acceptable artifacts. Images that were discarded in this phase have been used as Validation Set. As second step, only images with resolution size equal or greater than a specific threshold (set to 1024x1024) were kept. In fact, some neural networks architecture requires as input images of specific size. Images that were discarded in this second phase have been included into the Test Set to validate the proposed GAN models' performance. Data analysis leads to almost half the total amount of initial images from 33000 to about 18000 samples.

2.2.2 Image Processing

The application of image processing techniques using the Python libraries *openCV* and *skimage* was necessary in order to prepare the images for use as inputs for GAN models. This involved a multi-step process involving multiple conditional loops.

Dark Corner Artifact Removal

The first step involved checking for the presence of Dark Corner artifacts. To do this, a border frame of the image with a selected thickness (20 pixels) was analyzed. If the mean intensity of the border frame was found to be below a predetermined threshold, it was considered to be near-black, indicating the presence of the black frame artifact. In this case, the objective was to delimit the useful skin areas by

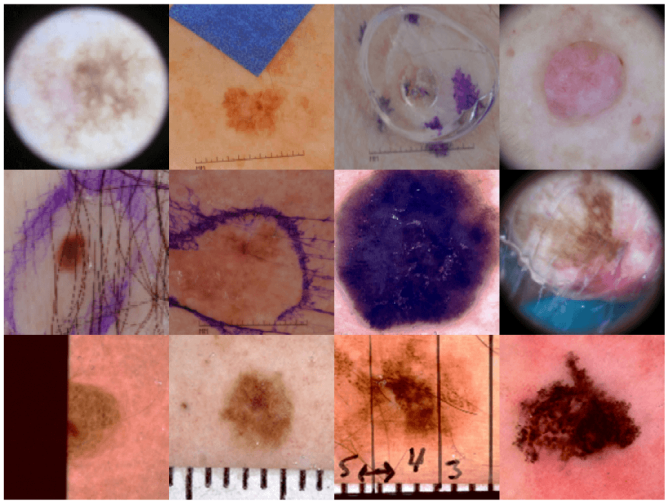


Figure 2.1: Examples of dermoscopy images affected from artifacts such as hair strands, ruler measures, air bubbles, and etc.

identifying their edges and reducing the dark frame areas. The following pipeline was applied: an image was loaded and read, and subsequently transformed from the RGB color space to grayscale. Otsu thresholding was applied to create a binary mask, representing the region of interest (ROI), by identifying the optimal threshold value that maximizes the variance between two classes of pixels. To refine the mask and remove any noise or small irrelevant regions, morphological operations were applied, specifically opening and closing. The contours of the mask were then calculated, and among them, the largest contour was selected. The bounding box of the selected contour was computed to acquire the coordinates of the ROI, which were then utilized to crop the original RGB image around the region of interest. These steps were crucial in identifying and extracting the ROI from the image, which was then used for further analysis.

Rectangular to Square Image Splitting

The second step involves checking the dimensions of the image. If the image is already squared (i.e., the height and width are equal) or even better, 1024x1024, it is saved without any further processing, except for a resize operation if necessary. If the image has a rectangular shape of a larger size resolution, it must be resized to the desired dimensions. In this case, the difference between the height and width is compared to a threshold (set to 1000). If the difference is larger than this threshold, the image is split into three sub-images and saved, as shown in Figure 2.3. Otherwise, a simple resize operation is performed on the complete image.
At the end of this second pipeline process, all images will have a dimension of

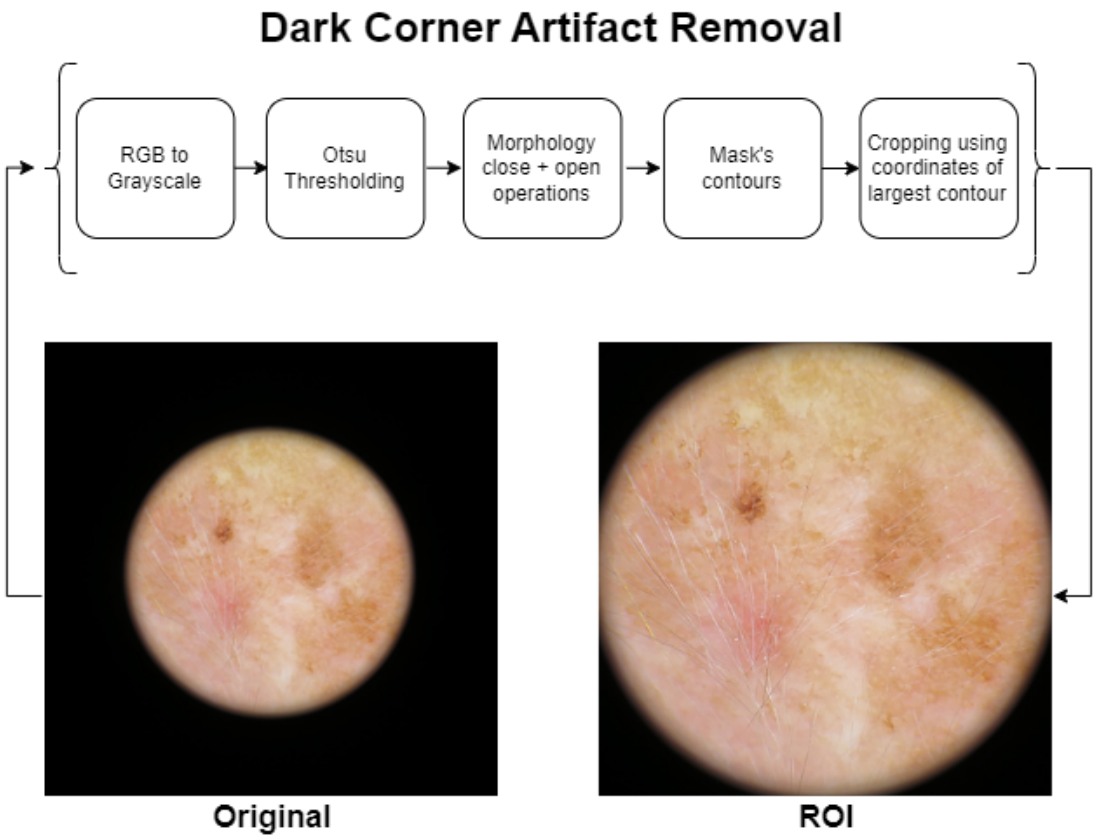


Figure 2.2: Dark Corner Artifact Removal: pipeline process.

1024x1024.

2.2.3 Data Augmentation

Deep neural networks often require a large amount of data to achieve good performance. Data augmentation is a technique that can be used to increase the number of available images during training by applying randomized alterations to the data, thereby increasing the variability of the dataset. This can help to reduce overfitting in deep learning models and improve performance on benchmark datasets by making the model more robust through the use of a larger amount of data during training. For this study, data augmentation was applied using the Python library *Albumentations* [42], which provides fast and flexible image augmentations by implementing a wide range of image transform operations for various computer vision tasks. Classic augmentation techniques such as affine transformations were carefully chosen and tested with appropriate parameters in order

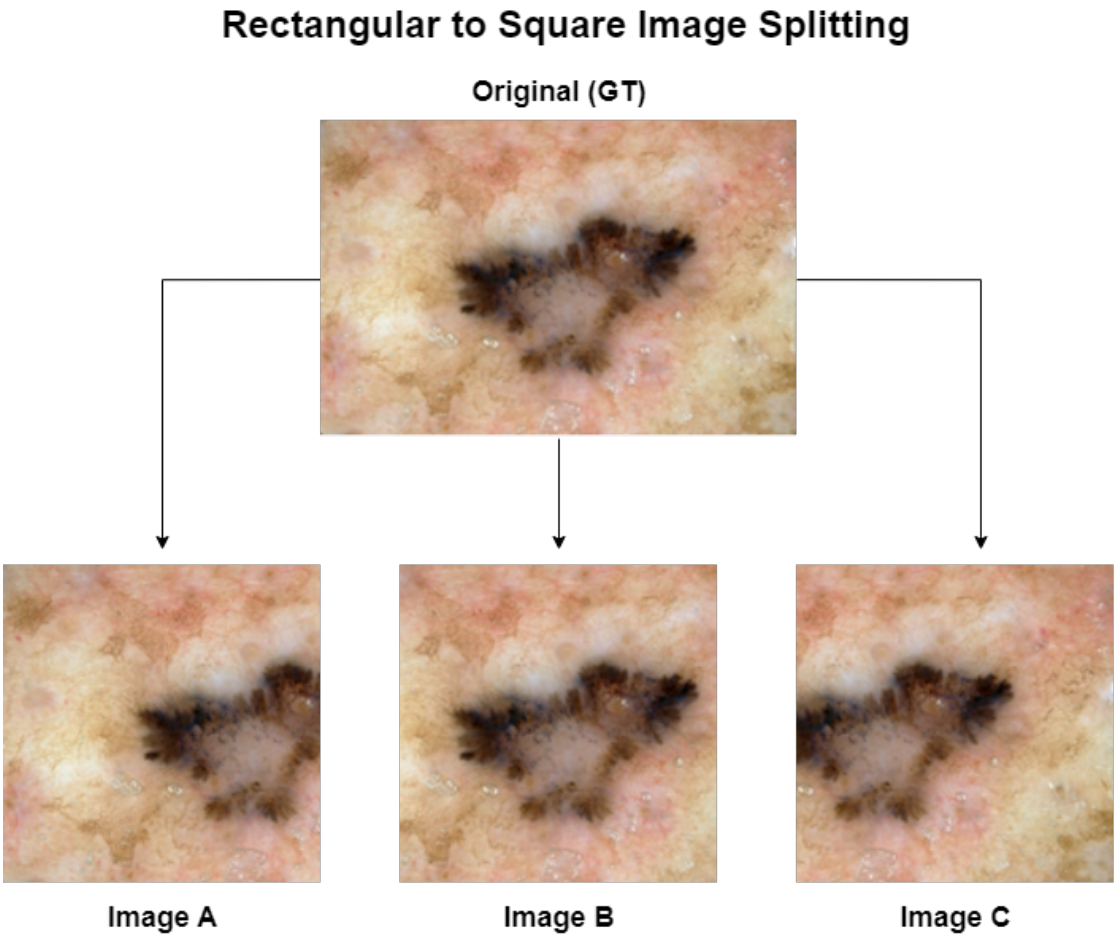


Figure 2.3: Rectangular to Square Image Splitting: pipeline process.

to preserve the characteristics of the skin lesions. In addition to these commonly used transforms, other operations such as grid distortion and elastic transform were lightly applied, as medical imaging often deals with non-rigid structures that may exhibit shape variations. Each transform in Albumentations has a parameter that sets the probability of applying that augmentation transform to the given image within the augmentation pipeline. The following data augmentation techniques were used:

- **Flipping:** due to the nature of dermoscopic images, flipping augmentation is particularly suitable. It was applied to one or both axes (vertical or horizontal). The flipping probability was set to 1, indicating that all images were subjected to flipping operations. Given the three possible flipping operations, namely vertical flip, horizontal flip, and a combination of both and the no flipping

transformation, each operation was assigned a probability of 25%. This ensured that each image had an equal chance of being flipped in any of the three ways or not flipped at all.

- **Scaling and shearing:** images of the same skin lesion are often taken multiple times with different zoom ratios and angles.
  To simulate similar conditions, a scaling operation was applied, randomly zooming in within the interval $[0.01\% - 0.05\%]$, while keeping the original aspect ratio. Additionally, a shearing operation was applied, slightly stretching the image within an interval range of [-5,5] degrees. The probability of both operations occurring was set to 30%. The probability of both operation to occur was set the 30%.

- **Brightness, Contrast and Gamma:** these ensure that the trained network is robust to different intensity differences, as images of the same skin lesion are often taken multiple times under different lighting conditions. Brightness and contrast were randomly applied together using a brightness limit of 0.1 and a contrast limit of 0.2, which are amount parameters that control the intensity of the operation transform. The Gamma transform was applied using a range of [60,140]. The probability of both transforms was set to 33%.

- **Grid Distortion and Elastic Transform:** these are spatial augmentations commonly applied to biomedical images [43]. They are similar to stretching algorithms, but more complex and with more freedom.
  Both geometric transforms were set to occur with a probability of 40%.

Furthermore, each of the four main transformations provided additional options that enabled the omission of the transformation altogether. This added an extra degree of heterogeneity and randomness to the dataset, providing the model with a wider range of variation in the training data, which in turn increases its generalizability and robustness. It is important to avoid augmentations that are unlikely to occur in real-world dermoscopic images. Moreover, every single augmentation transform was carefully tested to optimize the range of parameters values. This is crucial in order to preserve the original anatomical content of the images.

2.2.4 Paired Dataset Generation

The task of Super Resolution Image Reconstruction necessitates a dataset comprising pairs of low-resolution (LR) and high-resolution (HR) images. In the present study, custom degradation and enhancement procedures were employed to transform existing images, including augmented ones, into LR and HR images, respectively.
Real-world images are commonly subject to unknown and complex degradation

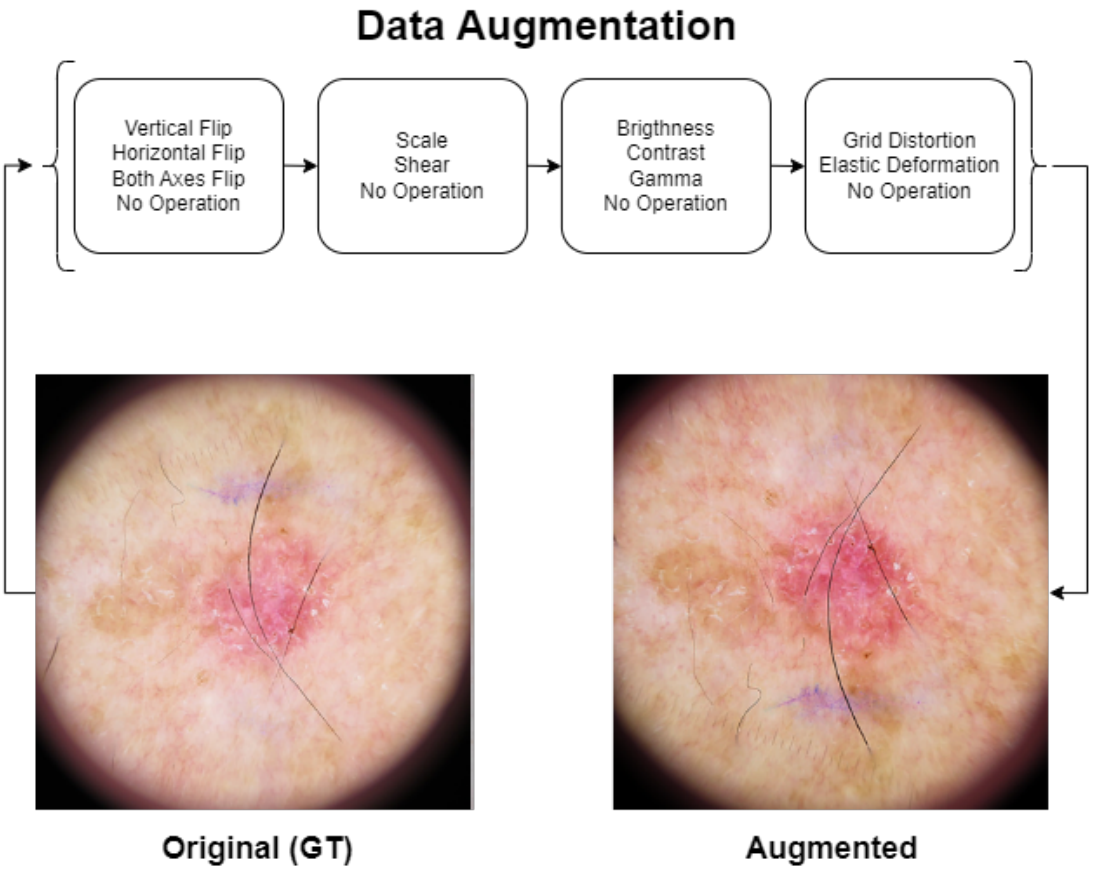


Figure 2.4: Data Augmentation: pipeline process.

processes, such as noise and compression artifacts. Consequently, models trained on datasets generated using manual degradation methods may not exhibit desirable performance in real-world scenarios. Blind super-resolution approaches aim to restore LR images degraded by unknown and complex processes while most classical SR methods assume a bicubic downsampling kernel and usually fail in real images.

Degradation

Existing methods can be broadly classified into explicit and implicit modeling based on the underlying degradation process [44]. Explicit models employ classical techniques such as blurring, noise addiction, downscaling, and JPEG compression, while implicit models utilize data distribution learning with Generative Adversarial Networks (GANs) to learn the degradation model. However, these methods are limited to the degradations observed in the training dataset and may not generalize

effectively to out-of-distribution images. In this study, the pipeline implemented in the Real-ESRGAN [44] was replicated with some customization: firsly a classic explicit method was elaborated, involving the following process:

- **Blur:** blur degradation was implemented as a convolution with a linear blur filter (kernel). Both isotropic and anisotropic Gaussian filters were used, as well as a Generalized Normal filter (Advanced) with randomly selected kernel parameters. In addition, a 2D sinc filter was also included to create ringing or overshoot artifacts. Since three different types of Blur filters were used, each was set to occur with a probability of 33.3% to ensures that the images were subjected to a diverse range of blur effects.

- **Noise:** two commonly-used noise types were used: additive Gaussian noise and Poisson noise. Additive Gaussian noise was used, which has a probability density function equal to that of the Gaussian distribution. The noise intensity is controlled by the standard deviation (i.e., sigma value) of the Gaussian distribution. When each channel of RGB images has independent sampled noise, the synthetic noise is color noise. To also synthesize gray noise, the same sampled noise was applied to all three channels. Poisson noise, also referred to as ISO in Albumentations, follows the Poisson distribution. It is usually used to approximately model the sensor noise caused by statistical quantum fluctuations, that is, variation in the number of photons sensed at a given exposure level. Poisson noise has an intensity proportional to the image intensity, and the noises at different pixels are independent. In this case, the three types of noise were set to be applied with a probability of 33.3% each.

- **Downscale:** it is an operation for synthesizing low-resolution images which decreases image quality by downscaling and upscaling back. There are several algorithms such as nearest-neighbor interpolation, area resize, bilinear interpolation, and bicubic interpolation. In order to include more diverse and complex downscaling effects, a random interpolation was performed from the above four choices, with random scale factor interval, each one with a probability of 25%.

- **JPEG Compression:** compression is a commonly used technique of lossy compression for digital images. Unpleasing block artifacts are usually introduced by the JPEG compression. The quality of compressed images is determined by a quality factor $q \in [0,100]$, where a lower q indicates a higher compression ratio and worse quality. The probability of compression to occur was set to 80%

A classic degradation pipeline typically involves convolving the image with a blur kernel k, simulating the effect of blur that may occur due to camera shake or

other factors. Next, a downscaling operation with a scale factor (d) is performed, decreasing the quality of the image and simulating the effect of low-resolution. The result of this process is the low-resolution image, which is further degraded by the addition of noise (n). Finally, the image is compressed using the JPEG algorithm, as this is a widely used method for image compression and commonly encountered in real-world images, resulting in the output image, y. However, to synthesize more practical degradations and mimic the real degradation generation process, the classical degradation model can be repeated one or two more times (m), as shown in the equation 2.1, with the addition of a random downscale and different parameters randomization. This approach increases the diversity of the degraded images, making the model more robust to real-world scenarios [44]. The number of repetitions is also based on the computational resources that are available, to reach a good balance between simplicity and effectiveness, and to ensure that the model is not overfitting on the training data.

$$y(x) = [((x \circledast k)_d + n)_{JPEG}]^m \tag{2.1}$$

Enhancement

On the other hand, high-resolution images were obtained through the use of a sharpening algorithm, which enhanced the sharpness and visual definition of texture edge contours. This was followed by the application of a light Gaussian smoothing filter to prevent the introduction of visible artifacts.
Eventually, the following generated paired image dataset is subsequently prepared for input into GAN models:

	Paired Dataset		**# Images**
Training Set	Low Resolution	High Resolution	36618
Validation Set	Low Resolution	High Resolution	3039

Table 2.2: LR-HR paired image dataset generated to train the model.

2.3 Network Architecture: Real-ESRGAN

The Real-ESRGAN [44] is an adapted version of the Enhanced Super-Resolution GAN (ESRGAN) model [45], which itself was an improved version of the SRGAN

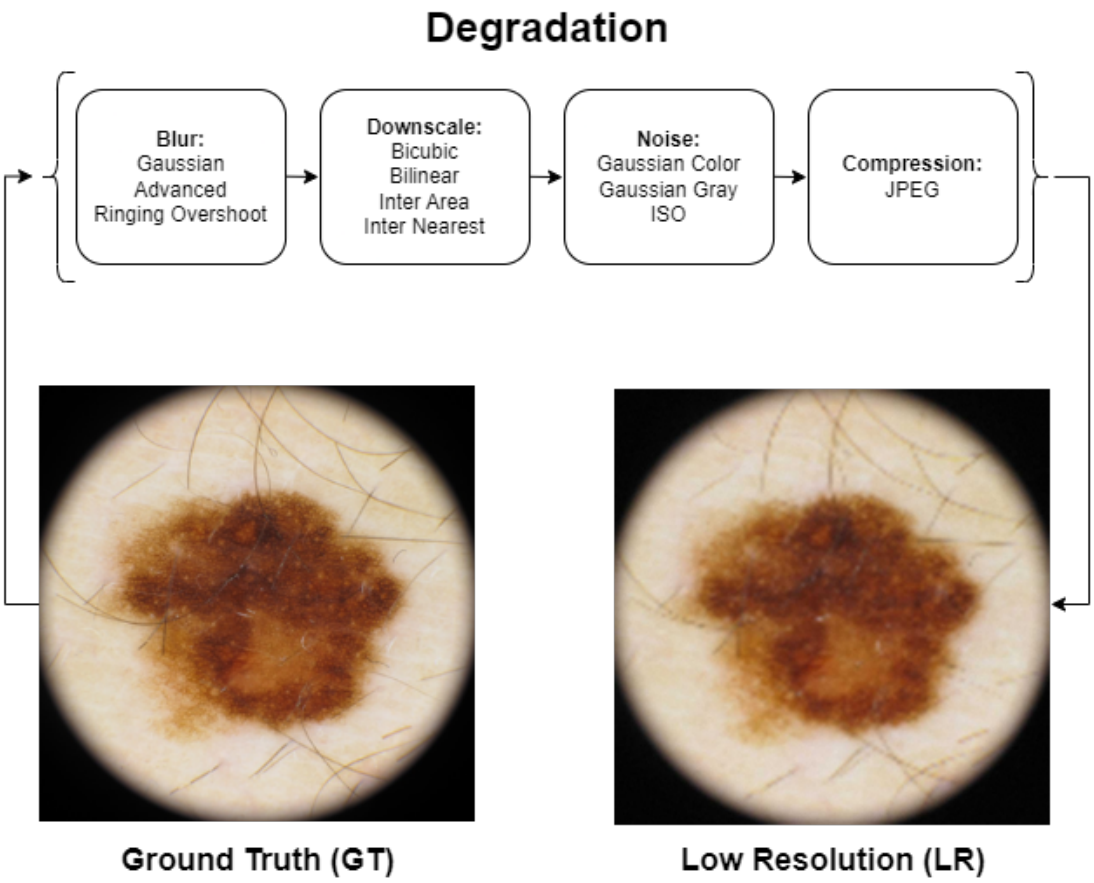


Figure 2.5: Degradation: pipeline process.

model. The goal of the Real-ESRGAN is to restore a greater number of real-world images and achieve better visual performance compared to previous works, making it more practical for use in real-world applications. To achieve this, the Real-ESRGAN model has improved upon certain aspects of the original SRGAN and ESRGAN network architecture, including:

- **Generator:** the generator of the Real-ESRGAN model has been modified in order to improve the quality of the recovered images compared to the SRGAN model. These modifications include replacing the original basic block with a Residual-in-Residual Dense Block (RRDB), which combines multi-level residual network and dense connections. This results in a higher capacity and easier training. Batch Normalization (BN) layers were removed, as they have been found to introduce artifacts. Removing BN layers also helps to improve generalization ability and reduce computational complexity and memory usage. In addition, residual scaling and smaller initialization were used to facilitate

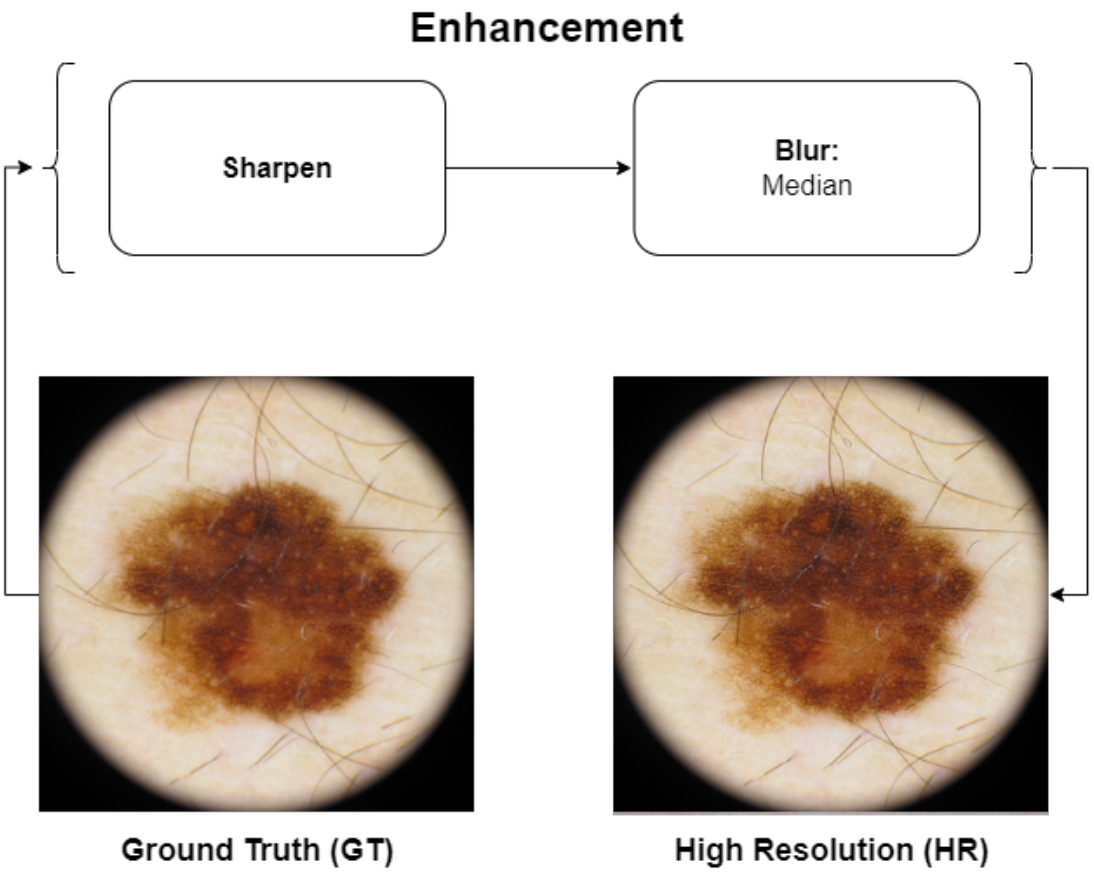


Figure 2.6: Enhancement: pipeline process

the training of a very deep network, as the residual architecture is easier to train when the initial parameter variance is smaller. The Real-ESRGAN maintains the same generator architecture as the ESRGAN.

- **Discriminator:** the discriminator of the Real-ESRGAN model has been modified in order to address a larger degradation space compared to the ESRGAN model. The original design of the discriminator, a VGG-style discriminator, has been improved to a U-Net design with skip connections. The U-Net structure provides detailed per-pixel feedback to the generator, which can improve local details, but also introduces unnatural textures and increases training instability. To address these issues, spectral normalization (SN) regularization is used to improve restored textures and stabilize the training dynamics. Using the U-Net discriminator with SN increases the discriminator's capability and stabilizes the training dynamics.

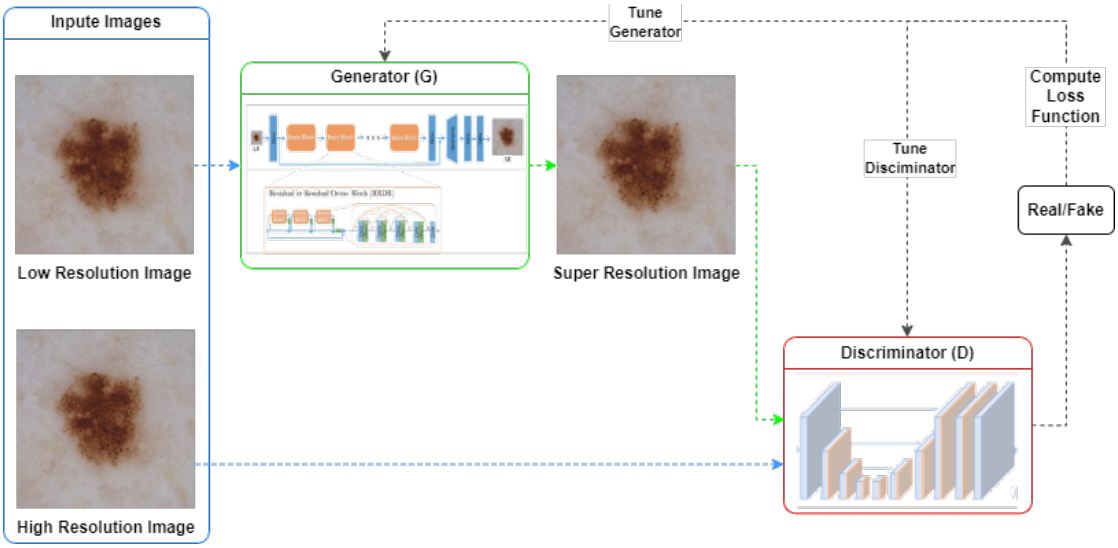


Figure 2.7: Real-ESRGAN scheme

- **Perceptual Loss:** in order to improve the effectiveness of the loss function, a modified perceptual loss was developed for the Real-ESRGAN model. This loss function uses the VGG features before activation, rather than after activation as in the SRGAN model. This modification has been found to produce sharper edges and more visually pleasing results, as it has been demonstrated empirically. Consequently, the total loss for the generator can be expressed as follows:

$$L_{RealESRGAN} = \gamma L_{percep} + \lambda L_{GAN} + \eta L_1 \tag{2.2}$$

where L_{percep} represents the Perceptual Loss and γ serves as the hyperparameter that regulates the weight of this loss. The adversarial loss function, L_{GAN}, is expressed as follows:

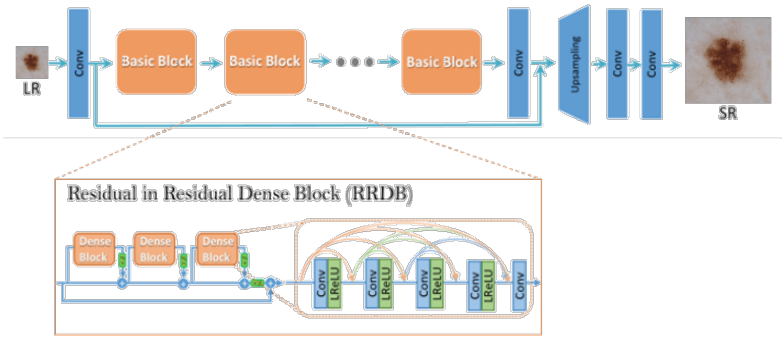
$$L_{GAN} = \log D(y) + 1 - \log D(G(x)) \tag{2.3}$$

L_{GAN} is controlled by the hyper-parameter λ. Finally, L_1 is the L1 normalization term which aims to produce outputs as close as possible to the target, with the hyper-parameter η regulating its weight.

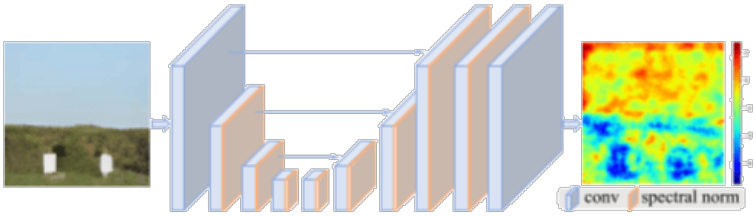
$$L_1 = ||y - G(x)||_1 \tag{2.4}$$

2.4 Training

Instead of training the Real-ESRGAN network from scratch, the transfer learning strategy has been applied to reduce the overall training time and costs [28]. The



(a) Real-ESRGAN Generator



(b) Real-ESRGAN Discriminator

Figure 2.8: Real-ESRGAN architecture.

concept of transfer learning allows models that are trained from scratch on general large-scale image datasets, in this study the ImageNet dataset [46], to be specialized for specific tasks by using a considerably smaller dataset that is problem-specific. In order to evaluate the performance of the network under different conditions, it was trained three times using varied configurations. Each training session involved the alteration of only one of the three main components of a deep neural network: the training data, the hyperparameters and the architecture. Specifically:

- **Version 0:** in the first training session, the model was subjected to training utilizing the initial paired dataset, that had been generated using a simplified pipeline of data augmentation and degradation processes. Specifically, the augmentation process excluded the distortion transformation, while the degradation pipeline only comprised downscaling and noise addition operations. Moreover, degradation was applied only once using a downscaling with a fixed scale factor of 4 and certain augmentation parameters were either fixed or adjusted within a restricted range. This version serves as a baseline for evaluating any future modifications made to the training dataset and hyper-parameters.

- **Version 1:** for the second training session, a new paired dataset was generated through a more complex augmentation and degradation pipeline, as detailed

in Chapters 2.2.3 and 2.2.4. For instance, the degradation process was applied twice, using a downscaling operation with a randomly selected scale factor in the interval [4,8]. Additionally, all feasible parameters were randomly adjusted within a wider yet acceptable range in the augmentation and degradation processes. The architecture and hyperparameters were maintained consistent with those utilized in the Version 0, in order to evaluate the impact of expanding the diversity, thus increasing the heterogeneity, of the training set.

- **Version 2:** in the third training session, the dataset utilized was consistent with that employed in the Version 1. However, modifications were made to the hyperparameters, such as the optimizer configurations. An examination was conducted to determine the appropriate number of iterations (epochs) and scaling criterion through experimentation with various learning rate schedules. The Adam optimization algorithm was replaced with AdamW, and additional optimization algorithms such as Stochastic Gradient Descent (SGD) were experimented. Additionally, modifications were made to the loss configurations and the hyperparameters that regulate their weights. Several experiments were performed in which different loss configurations were implemented. In particular, the 'vanilla' L_{GAN} Loss was substituted with the 'WGAN' Loss and the 'LS' Loss, and the $L1_{GAN}$ Loss was replaced by the 'MSE (L2)' Loss and the 'Charbonnier' Loss. It is worth noting that the $L_{perceptual}$ Loss remained unchanged across all experiments.

Throughout all training sessions, the architecture configurations were kept the same, consisting of a generator formed by 23 Residual-in-Residual Dense Blocks (RRDBs) and a U-Net serving as the discriminator. The most optimal training session out of the three listed was conducted for approximately 100000 iterations, employing a batch size of 1 and utilizing the Adam optimization algorithm with an initial learning rate of 10^{-3}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Additionally, a learning rate scheduler was implemented, which decreased the learning rate by a factor of 0.5 after every 25000 iterations, while retaining the original loss configurations used in the Real-ESRGAN original implementation. All the experiments have been conducted on a NVIDIA RTX 3090 GPU (24GB VRAM).

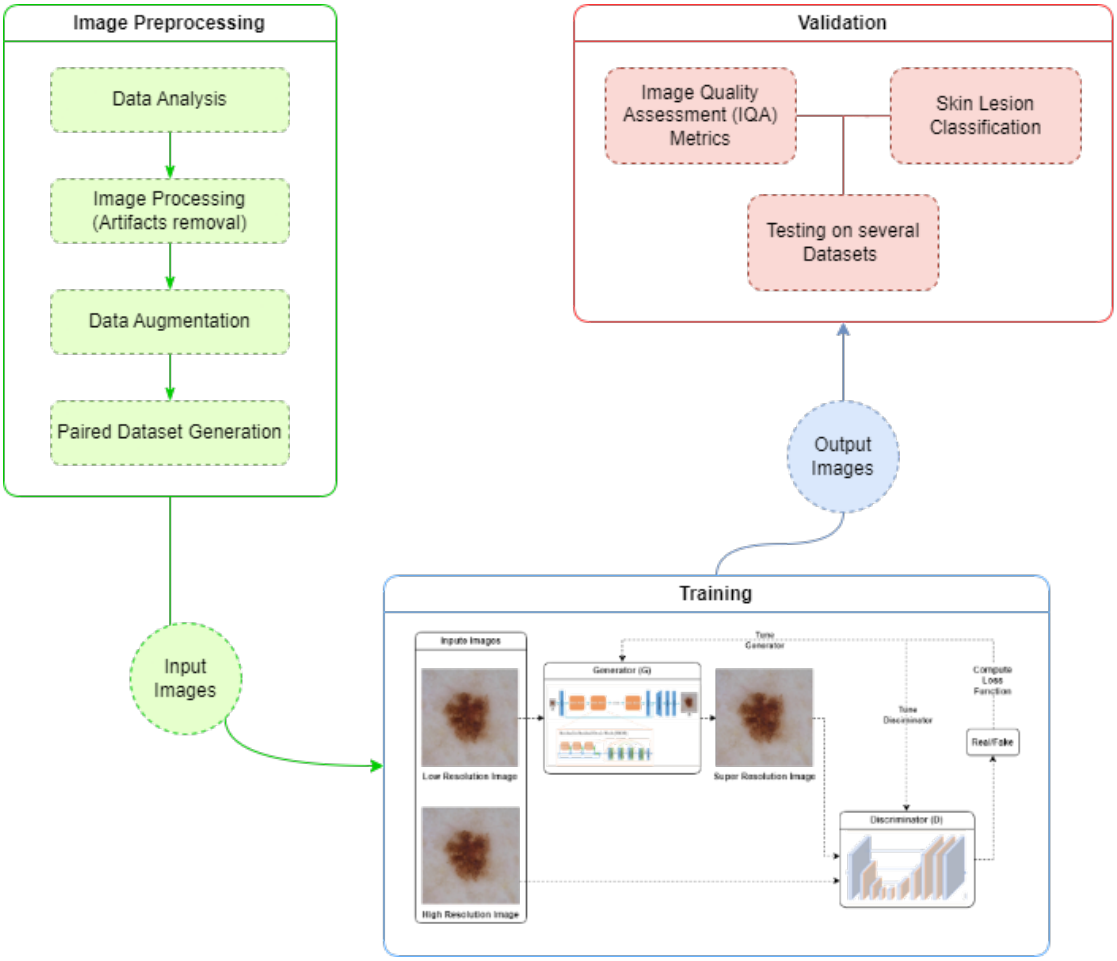


Figure 2.9: Summarization scheme of the proposed Real-ESRGAN based approach main steps: (a): analyzing, processing, augmenting and generating LR-HR pair dataset. (b): employing the Real-ESRGAN to generate SR images. (c): validation and evaluation through Testing on several datasets, IQA metrics and skin lesion classification.

Chapter 3

Results

3.1 Evaluation Methods

3.1.1 GANs Metrics

As the utilization of generative adversarial networks (GANs) in various real-world applications has increased with the advancement of GAN architectures, various qualitative and quantitative metrics have been proposed for evaluating GAN performance [47]. However, a universally accepted benchmark metric for evaluating the comprehensive performance of a GAN architecture has not yet been established. Despite this, certain essential and desirable characteristics have been identified for GAN evaluation metrics, which enable the measurement of GAN performance and facilitate comparison. These characteristics include: the preference for models to generate samples that are markedly distinguishable from real ones, sensitivity to overfitting, clearly defined boundary values, responsiveness to image distortions and transformations, consistency with human perceptual judgments and evaluations of models, and low sample and computational complexity. Some of the most utilized metrics for measuring performance of GANs are:

- **Inception Score (IS):** measures the quality of generated images by evaluating their diversity and their perceived realism. It is based on the idea that a model that generates high-quality images will also produce a diverse range of images. The IS is calculated using the Inception v3 model, which is a pre-trained image classification model. The generated images are fed into the Inception v3 model, and the model's output is used to estimate the entropy of the generated images. The IS is then calculated as the average KL divergence between the predicted and true class distributions, where the true class distribution is approximated by sampling a large number of images from the training set. One advantage of the IS is that it is relatively easy to implement and compute.

It also provides a single scalar score that can be used to compare different GAN models. However, the IS has several limitations. First, it relies on the Inception v3 model, which may not always be the most appropriate choice for evaluating GAN models. Second, the IS does not take into account the overall quality of the generated images, only their diversity and perceived realism. Finally, the IS is sensitive to the size of the generated image dataset, which can make it difficult to compare GAN models trained on different datasets [47].

- **Frechet Inception Distance (FID):** is a metric that measures the distance between the distributions of real and generated images. It is based on the idea that if the distributions of real and generated images are similar, then the generated images are of high quality. The FID is calculated by first extracting the activations of the Inception v3 model for both the real and generated images. The activations are then used to estimate the means and covariances of the real and generated image distributions. The FID is then calculated as the square root of the sum of the squared differences between the means and the Frobenius norm of the difference between the covariances. A lower FID value indicates a smaller distance between the two distributions. FID is known for its strong discriminative capabilities, robustness, and computational efficiency. However, it is based on the assumption that the features of the data follow a Gaussian distribution, which is not always the case [47].

3.1.2 Image Quality Assessment (IQA)

As the use of digital imaging and communication technologies continues to expand and become more diverse, it becomes increasingly important to accurately assess the quality of images. The goal of image quality assessment (IQA) is to develop metrics that can automatically predict the visual quality of an image based on its perceived attributes by human viewers. This process can be divided into subjective methods, which rely on human perception, and objective methods, which use computational models to predict image quality. Subjective methods more closely align with human perception, but can be inconvenient, time-consuming, and expensive. Therefore, objective methods are currently the most widely used approach to IQA. However, there can be significant discrepancies between the results of subjective and objective methods due to the difficulties in accurately capturing human visual perception with objective methods [48].
Image quality can be impaired by various distortions during image acquisition and processing, such as noise, blur, ringing, and compression artifacts. A valuable quality metric should closely align with the subjective perception of quality by a human observer. Quality metrics can also be used to track unperceived errors as they propagate through an image processing pipeline and can be used to compare

the performance of different image processing algorithms. If a reference image without distortion is available, it can be used to directly compare the target and reference images by using Full-Reference (FR) metrics. For example, when assessing the quality of compressed images, an uncompressed version of the same image can be used as a reference. If a reference image is not available, No-Reference (NR) image quality metrics can be used instead. These metrics calculate quality scores based on expected image statistics. The computation of the following metrics was performed utilizing functions implemented within *Matlab*.

Full Reference (FR) Metrics

Full-reference algorithms compare the input image against a pristine reference image with no distortion [49].

- **Peak Signal-to-Noise Ratio (PSNR):** it is a widely utilized metric for evaluating the reconstruction quality of lossy image transformation techniques (e.g., image compression), by comparing the quality of a generated image to a reference image. PSNR is computed using the maximum possible pixel value and the Mean Squared Error (MSE) between the images. Given a reference image, denoted as GT, and a reconstructed image, denoted as SR, the MSE and PSNR (in dB) between GT and SR are defined as follows:

$$MSE(x,y) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-y_{ij})^2 \tag{3.1}$$

$$PSNR(x,y) = 10 * \log_{10}\frac{\{\max[\max(x),\max(y)]\}^2}{MSE} \tag{3.2}$$

  In general cases using 8-bit image, the maximum pixel value is equal to 255 and PSNR typical values range from 20 to 40, with higher values indicating a higher quality image. However, PSNR only takes into account the difference in pixel values at the same positions, rather than considering human visual perception, which can result in inadequate representation of the quality of super-resolved images in real-world scenarios where human perception is of greater importance. Despite this limitation, PSNR remains the most widely used evaluation criterion for super-resolution models due to its ability to facilitate comparisons with previous work and the lack of entirely accurate perceptual metrics.

- **Structural Similarity Index Measure (SSIM):** is a method of quantifying the similarity between two images by attempting to model the perceived change in the structural information of the image, and has been widely used in image processing and computer vision due to its ability to provide a more accurate

representation of the perceived quality of images compared to other commonly used metrics such as peak signal-to-noise ratio (PSNR). This approach is based on the idea that the human visual system (HVS) is highly adept at extracting structural information from the visual field, and thus SSIM is proposed as a method of measuring the structural similarity between images by taking into account three relatively independent comparisons: luminance, contrast, and structure. SSIM can be expressed through these three terms as:

$$SSIM(x,y) = \left[(l(x,y))^{\alpha} (c(x,y))^{\beta} (s(x,y))^{\gamma} \right] \tag{3.3}$$

Here, l is the luminance (used to compare the brightness between two images), c is the contrast (used to differ the ranges between the brightest and darkest region of two images) and s is the structure (used to compare the local luminance pattern between two images to find the similarity and dissimilarity of the images) and α, β and γ are the positive constants. Again luminance, contrast and structure of an image can be expressed separately as:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$
$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$
$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

If $\alpha = \beta = \gamma = 1$, then the index is simplified as the following form:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.4}$$

where x and y are the two images being compared, μ_x and μ_y are the means of the images, σ_x^2 and σ_y^2 are the variances of the images, and σ_{x*y} is the covariance between the images. $C1 = (k_1 \cdot L)^2$ and $C2 = (k_2 \cdot L)^2$ are constants for avoiding instability, with $k1 \ll 1$ and $k2 \ll 1$ being small constants, and L is the maximum possible pixel value. The SSIM value varies between -1 and 1, where a value of 1 shows perfect similarity.
In addition, due to the possible unevenly distribution of image statistical features or distortions, assessing image quality locally may be more reliable than applying it globally. Thus mean structural similarity (MSSIM) can be used for assessing SSIM locally. Specifically, it splits the images into multiple windows, assesses the SSIM of each window, and finally averages them as the final MSSIM.

- **Multiscale Structural Similarity Index Measure (MS-SSIM):** is a multiscale version of SSIM that allows for more flexibility in incorporating

image resolution and viewing conditions than a single scale approach. MS-SSIM ranges between 0 (low similarity) and 1 (high similarity). It is an advanced version of SSIM which evaluates various structural similarity images at different image scale.

- **Feature Similarity Index Measure (FSIM):** is a technique for evaluating the similarity between two images based on their Phase Congruency (PC) and Gradient Magnitude (GM). PC is a method of detecting image features that is resistant to changes in lighting and contrast, and can also identify additional features by analyzing the image in the frequency domain. GM, on the other hand, involves calculating image gradients using convolution masks. The gradient magnitude of an image can be defined as the square root of the sum of the squares of the horizontal and vertical gradients of the image. The phase congruency of f1 and f2 can be represented by PC1 and PC2, respectively, and the gradient magnitude maps extracted from the two images can be denoted by G1 and G2. The FSIM can be calculated based on PC1, PC2, G1, and G2. First, the similarity between the two images can be calculated using their phase congruency maps as follows:

$$S_{pc} = \frac{2PC_1 \cdot PC_2}{PC_1^2 + PC_2^2 + T_1} \tag{3.5}$$

where T_1 is a positive constant that increases the stability of S_{pc}. Similarly, the similarity between the two images can be calculated using their gradient magnitude maps as follows:

$$S_g = \frac{2G_1 \cdot G_2 + T_2}{G_1^2 + G_2^2 + T_2} \tag{3.6}$$

Where T_2 is a positive constant that depends on the dynamic range of gradient magnitude values. Finally, the overall similarity between the two images, SSIM, can be calculated by combining $\S_{pc}$ and S_g as follows:

$$FSIM = (S_{pc})^{\alpha} \cdot (S_g)^{\beta} \tag{3.7}$$

where α and β are parameters used to adjust the relative importance of PC and GM features, usually set to 1 for convenience. The resulting value of SSIM ranges from 0 to 1, with higher values indicating greater similarity between the images. While FSIM may be more effective at capturing human visual perception than other image quality assessment methods such as PSNR and SSIM, these latter methods are still more commonly used due to their historical prevalence.

Non Reference (NR) Metrics

In many situations, it is not possible to access the original or reference version of an image, making it difficult to use full reference methods for estimating the quality of a processed image. In these cases, efficient NR-IQA algorithms, also known as blind or no-reference image quality assessment (B/NR-IQA) algorithms, are designed to evaluate the quality of an image in an objective and reproducible way, even when a reference image is not available [48]. The BRISQUE and NIQE algorithms are both able to efficiently calculate the quality score of an image once the model has been trained. In contrast, the PIQE is less computationally efficient, but it has the added advantage of providing local measures of quality in addition to a global quality score. Overall, No-Reference quality metrics tend to outperform Full-Reference metrics in terms of their agreement with subjective human quality scores. This means that no-reference metrics are able to more accurately predict how humans perceive the quality of an image. As a result, no-reference metrics are widely used in a variety of applications for objectively and reproducibly evaluating the quality of images [50].

- **Natural Image Quality Evaluator (NIQE):** is based on models trained upon a database of pristine images to compute a quality score and it is used to measure the quality of images with arbitrary distortion. NIQE is opinion-unaware, and does not use subjective quality scores [48].
  It utilizes natural scene statistics (NSS) as predictable statistical features in their training process. NSS features, which are based on normalized luminance coefficients in the spatial domain and modeled as a multidimensional Gaussian distribution, are used to identify distortions in the image. To determine the NIQE score of images, Matlab niqe function was utilized, to extract natural scene statistics (NSS) features from statistically significant blocks in the distorted image. The function then fits a multivariate Gaussian distribution to these features and calculates the image quality score as the distance between this distribution and the distribution fitted to a reference image.

- **Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE):** is designed to evaluate the perceived quality of images that have undergone some kind of spatial processing or distortion. In order to accomplish this, a BRISQUE model is trained on a database of images with known distortions. As a result, BRISQUE is limited in its ability to evaluate the quality of images with different types of distortions. BRISQUE, unlike NIQE, is opinion-aware, meaning that subjective quality scores are provided alongside the training images. This allows the model to be trained on data that reflects human perceptions of image quality [50]. In terms of its training process, BRISQUE shares some similarities with the NIQE. However, it differs in how it utilizes

natural scene statistics (NSS) features to compute a quality score. Specifically, BRISQUE extracts NSS features from the distorted image and uses Support Vector Regression (SVR) to predict a quality score.

- **Perception based Image Quality Evaluator (PIQE):** is also an opinion-unaware and unsupervised algorithm, which means it does not require a trained model. PIQE is able to measure the quality of images with arbitrary distortion and has been shown to perform similarly to the Natural Image Quality Evaluator (NIQE) in many cases. To compute a quality score, PIQE estimates block-wise distortion and measures the local variance of perceptibly distorted blocks [51]. This approach allows PIQE to provide a detailed analysis of the quality of an image at a local level, in addition to a global quality score.

3.1.3 Texture Analysis

Color and texture are the two main types of characteristics used by dermatologists to differentiate between skin melanocytic patterns. Dermoscopic structures, such as reticular patterns and streaks, can be described using texture features, as these markers represent the spatial intensities in an image and allow for the identification of different shapes. Texture analysis methods can be divided into several categories: statistical, structural, model-based, and signal processing-based. Both texture and color are widely recognized as important factors in image analysis, and numerous approaches for texture analysis have been successful in image processing [52]. To conduct color texture analysis, skin images were converted to the YCbCr color space, extracting textural information from the luminance plane and chrominance features, which allows for a clear separation between texture and color features. In this study, texture analysis was performed on the Novara GQ Test Set using the *Pyradiomics* [53] library in Python. To do so, region of interest (ROI) segmentations, around the lesion, were manually created using the *ImageJ* software, and all available features were measured. Finally, only a few features that were considered most important for the study's application were selected.

Neighbouring Gray Tone Difference Matrix Features (NGTDM)

The Neighbouring Gray Tone Difference Matrix (NGTDM) is a matrix that quantifies the difference between a particular gray value and the average gray value of its neighbors within a certain distance δ [53]. Some extracted features are:

- **Coarseness:** a measure of the average difference between the central pixel and its neighbors, and it is an indicator of the spatial rate of change. A higher value of Coarseness corresponds to a lower spatial change rate and a locally more uniform texture

- **Contrast:** is a measure of the intensity change in the image and is also dependent on the overall gray level dynamic range. An image with a high contrast has a large range of gray levels and large changes between pixels and their neighbors.
- **Complexity:** reflects the number of primitive components in an image, with a higher value indicating a non-uniform image with many rapid changes in gray level intensity.
- **Strength:** is a measure of the primitives in an image, with a higher value indicating that the primitives are easily defined and visible, i.e. an image with slow intensity changes but large coarse differences in gray level intensities.

3.1.4 Task-based Evaluation

Another effective way to assess the quality of image reconstruction in super-resolution (SR) models is to evaluate the performance of other vision tasks using the original and reconstructed high-resolution (HR) images as input. This involves feeding the images into a trained model and comparing the impact on prediction performance. The vision tasks used for evaluation may include object recognition, image classification, and image segmentation, among others. This approach takes into account the fact that SR models can often improve the performance of other vision tasks [33].

Skin Lesion Classification

In order to evaluate the performance of the Real-ESRGAN, a deep learning network was implemented for the purpose of classifying skin lesions. The previous Test Set created from ISIC Archive Dataset was utilized as the initial dataset for the classification task. This dataset contained 7 classes, but was imbalanced, with insufficient instances of the VASC and DK classes in comparison to the remaining 5 classes. As a result, these two classes were removed. Data augmentation was then applied, in order to balance the number of images per label for the training phase. Finally, the dataset was partitioned into training, validation, and test sets in an 75%, 10%, 15% proportion, respectively: The ResNet50 architecture [30, 28], was selected for this task, involving a transfer learning approach by utilizing a pre-trained model on the large ImageNet [46] dataset and adapting it to a new task using a custom dataset. This approach saved time and improved performance in comparison to training a model from scratch. The deep learning network was trained using an input size of 512x512x3 pixels, to preserve as much information as possible, given the medical nature of the content. The batch size was set to 64 and a learning rate of $10^{-}3$ was initially used, which was decreased by a factor

ISIC Test Set				
Label	Training Set	Validation Set	Test Set	Total
AKIEC	921	103	181	1025
BCC	937	105	185	1227
KL	971	108	191	1270
ML	961	107	189	1257
NV	946	106	186	1238
Total	4,736	529	932	6,198

Table 3.1: The ISIC Test Set was further divided into new Training, Validation, and Test sets to be used in the classification tasks.

of 0.1 every 90 epochs using a scheduler. The Adam optimizer and Focal Loss were chosen as the optimization and loss function due to their efficiency and low computational cost, respectively. The training process was completed over a total of 360 epochs.
To evaluate the performance of the classifier, the following metrics were utilized: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's performance, including its ability to correctly identify positive cases (recall) and avoid false positives (precision). The F1-score is a harmonic mean of precision and recall, providing a balance between the two. These metrics were calculated using the following equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.8}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.10}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.11}$$

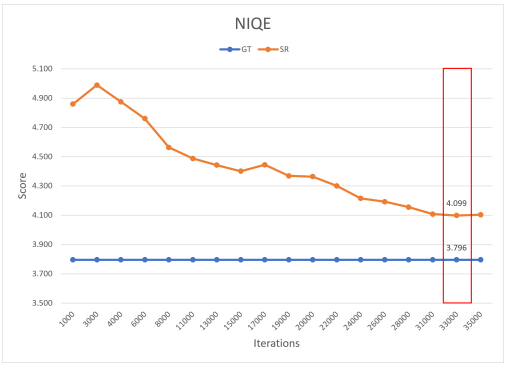
The term True Positive (TP) refers to a situation where the model's prediction of the positive class is in accordance with the actual condition. As an illustration, in the context of a medical diagnostic model, a True Positive result would signify that the model has accurately identified the presence of a disease in a patient who indeed has the disease. Conversely, True Negative (TN) refers to the scenario where the model's prediction of the negative class aligns with the actual condition. In the previously mentioned medical diagnostic example, a True Negative result would indicate that the model has correctly determined the absence of the disease in a healthy patient. False Positive (FP) refers to a situation where the model's

prediction of the positive class is inconsistent with the actual condition. In the medical diagnostic example, a False Positive result would signify that the model has erroneously diagnosed the presence of the disease in a healthy patient. False Negative (FN) refers to a scenario where the model's prediction of the negative class contradicts the actual condition. In the medical diagnostic context, a False Negative result would imply that the model has failed to detect the presence of the disease in a patient who indeed has the disease.

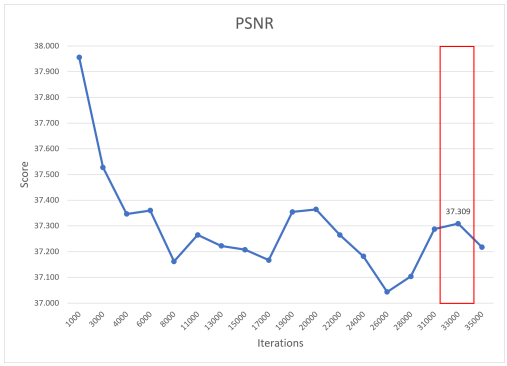
3.2 Results Analysis

3.2.1 Epoch Choosing Criteria

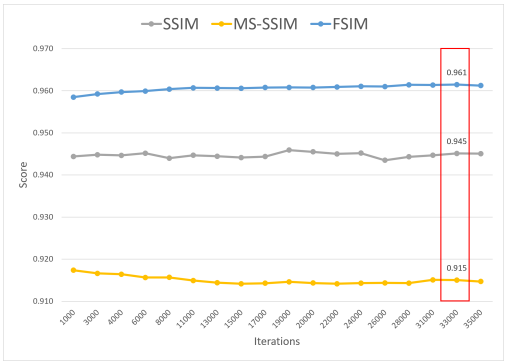
During the training phase of each model, the images within the validation set were repeatedly processed and restored to their super-resolution version utilizing the current model weights, at an interval of 5000 iterations, with the exception of the training Version 1, during which the model was saved every 2000 iterations. The model weights and corresponding super-resolution images were preserved at each evaluation point for further analysis. Upon the completion of the training session, the performance of each model was evaluated using the image quality metrics outlined in Chapter 3.1.2 of the study. To determine the optimal iteration from which to select the corresponding model weights for eventual inference, a specific focus was placed on the No-Reference metric known as NIQE, due to its effectiveness and the fact that it does not require the availability of a High Resolution reference image. In instances of uncertainty resulting from equivalent or similar NIQE metrics, the Full-Reference metrics scores of SSIM and FSIM were also taken into consideration, choosing the model weights belonging to the iteration with the higher score of SSIM and FSIM. Therefore, an hierarchical evaluation algorithm approach was implemented to enable the identification of the model iteration that demonstrated the highest level of image quality performance for the given task. Typically, the latter iterations of the training process exhibited the best metric scores, as the model's proficiency at the given tasks generally improves with increased training.
As previously mentioned in section 2.4, three different models were evaluated in order to identify the most effective dataset Preprocessing techniques and hyperparameter configurations for the Real-ESRGAN architecture to use in the task of reconstructing high-resolution skin lesion images from low-quality ones. In the case of training Version 0, the weights model belonging to the final iteration was chosen as the best, as a thorough examination of the training log information revealed that the PSNR and SSIM metrics were superior to those of previous iterations. As depicted in Figure 3.1, the model iteration 33000 was selected as the optimal model in the training Version 1 based on its superior performance in terms of the primary metric,



(a) NIQE



(b) PSNR

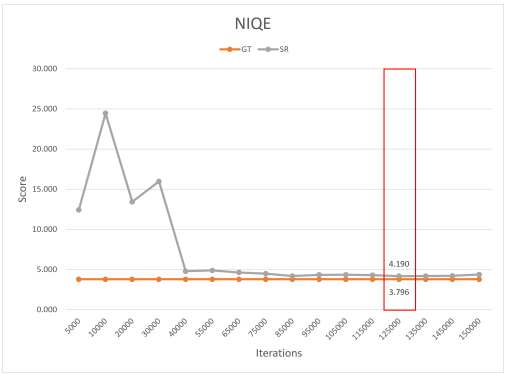


(c) SSIM, FSIM, MS-SSIM

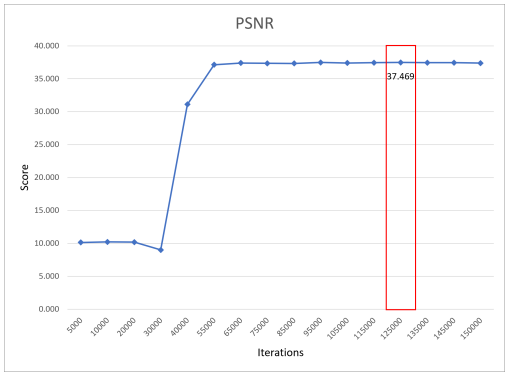
Figure 3.1: Epoch choosing criteria in the Training Version 1.

NIQE, and secondary metrics, SSIM and FSIM. Despite the fact that the PSNR value for this model was not the highest among the candidates, it was not given as much weight in the selection process due to its lower reliability as an IQA metric. However, it was still calculated as it holds historical and literature significance as an IQA metric.

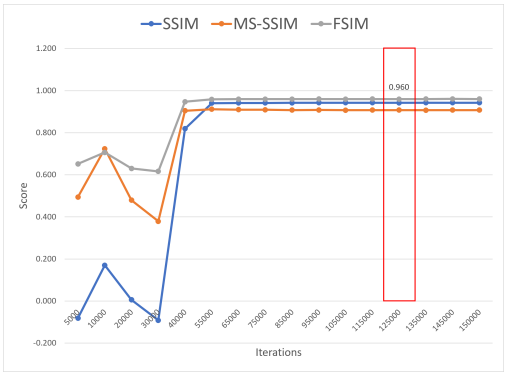
As illustrated in Figure 3.2, the model iteration 125000 was determined as the optimal model in training Version 2 utilizing the same criteria as in the previous training version. However, upon examination of the figure, it can be observed that the model's performance begins to converge around iteration 700000. This suggests that it may have been a more efficient strategy to terminate the training process prior to this point, as a model within that range of iterations would likely have performed comparably to the ultimately selected model, thus avoiding further resource and time expenditure. Ultimately, the three best models based on the training Versions 0, 1, and 2 were selected for use in restoring all images of the available Test Sets.



(a) NIQE



(b) PSNR



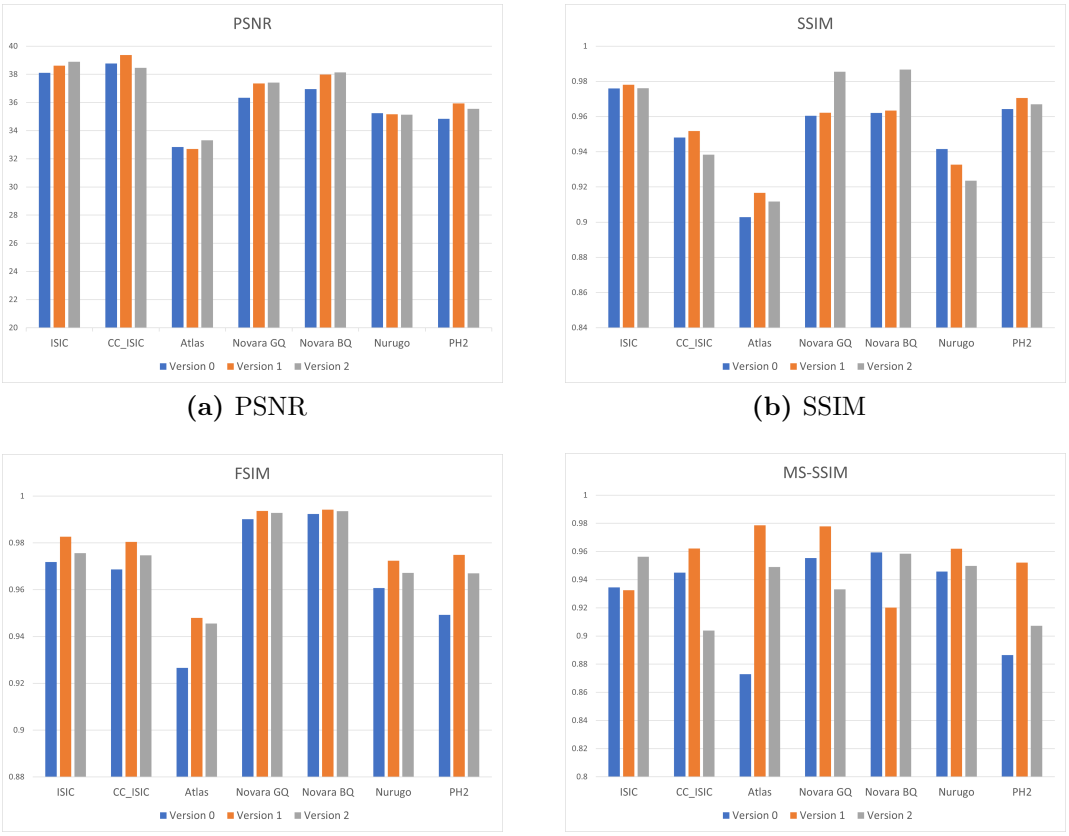
(c) SSIM, FSIM, MS-SSIM

Figure 3.2: Epoch choosing criteria in the Training Version 2.

3.2.2 IQA Metrics on Test Sets

Once the images of all available test sets were restored, two versions of each Test Set were obtained: the ground truth version, which contains the original images, and the super-resolution version, which contains the restored images obtained by restoring the ground truth image using the trained GAN model. This process was repeated for each of the three model obtained respectively from Training version 0, 1 and 2.
Subsequently, Image Quality Assessment metrics of all test sets were measured in order to obtain the metric scores for each image within each individual Test Set. The resulting values were then analyzed using statistics such as the mean score, standard deviation, and 5% - 95% percentile values. Moreover, these values were also graphically represented using histograms, as shown in Figure 3.3 for the FR metrics, and Figure 3.4 corncerning the NR metrics, to facilitate comparison and analysis. Other than comparing all the test sets results per model version, a direct comparison between ISIC and Color-Constancy ISIC Test Set was done to assess

what influence can bring, color standardization of skin images, in terms of quality and how this can affect the restoration task.

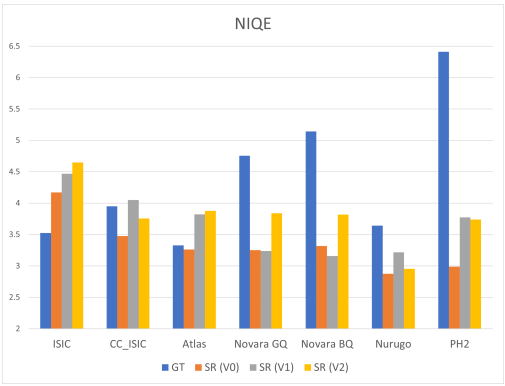


(a) PSNR **(b)** SSIM

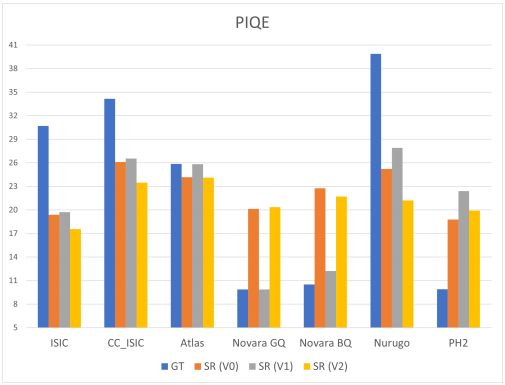
(c) FSIM **(d)** MS-SSIM

Figure 3.3: Image Quality Assessment Full-Reference (IQA-FR) metrics on available Test Sets.

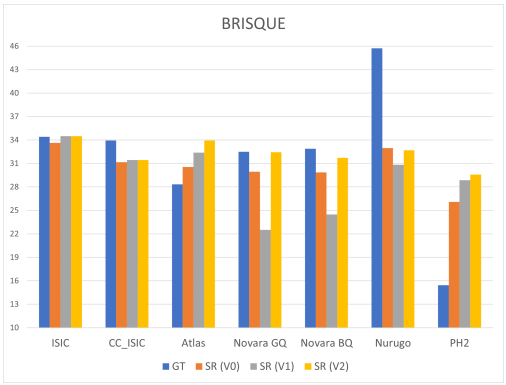
The results of the Full-Reference metrics indicate that utilizing a more diverse dataset and a more sophisticated degradation pipeline improves the performance of the GAN in restoring images from their low-resolution counterparts. Thus, it can be inferred that the model generated using the settings of Training Version 1 is generally superior. However, the results obtained from the No-Reference metrics are somewhat divergent and inconsistent both within each metric and between them. Nonetheless, it can be stated that the cases where restored Test Set images (SR) exhibit averagely superior performance in terms of NR IQA metrics when compared to the original images (GT), are greater than the ones with opposite situation (e.g., both the ISIC Test Sets and Atlas when referring to NIQE, or the Atlas and PH2 when referring to BRISQUE).



(a) NIQE



(b) PIQE



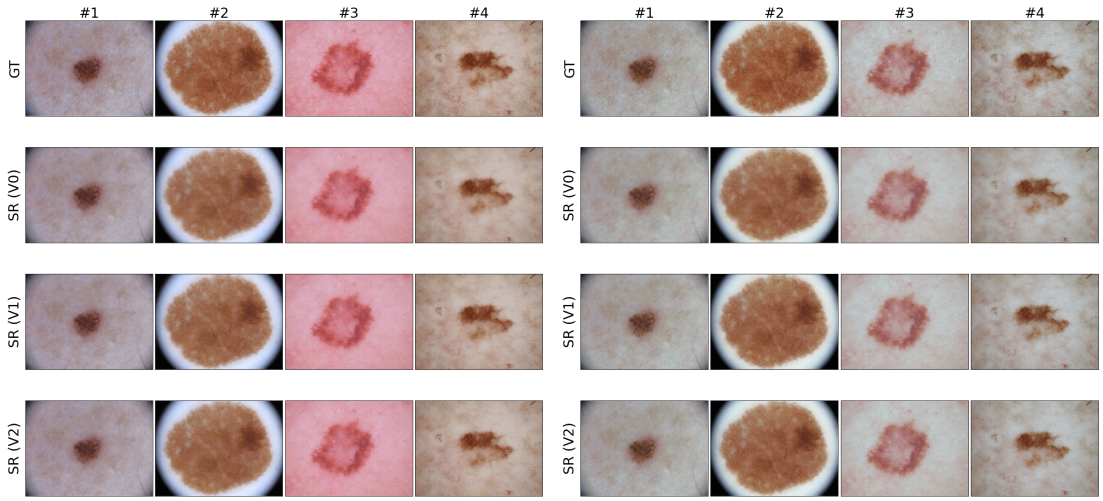
(c) BRISQUE

Figure 3.4: Image Quality Assessment No-Reference (IQA-NR) metrics on available Test Sets.

In terms of the comparison between the ISIC original and the color-standardized Test Set (CC ISIC), the Full-Reference metrics reveal a similar performance in the image restoration process for both datasets, with the exception of SSIM which shows lower performance for the Color-Constancy normalized dataset. On the other hand, the No-Reference metrics suggest an overall slightly better quality of restoration for both NIQE and BRISQUE, but a worse performance for the PIQE metric. Drawing from the obtained outcomes, it can be asserted that Real-ESRGAN demonstrates effectiveness in both situations. Therefore, it can be concluded that the color normalization technique does not negatively impact the task of reconstructing images. Rather, the color standardization and restoration operations executed by DermoCC-GAN and Real-ESRGAN, respectively, can be combined to produce higher quality images.

3.2.3 Visual Assessments on Test Sets

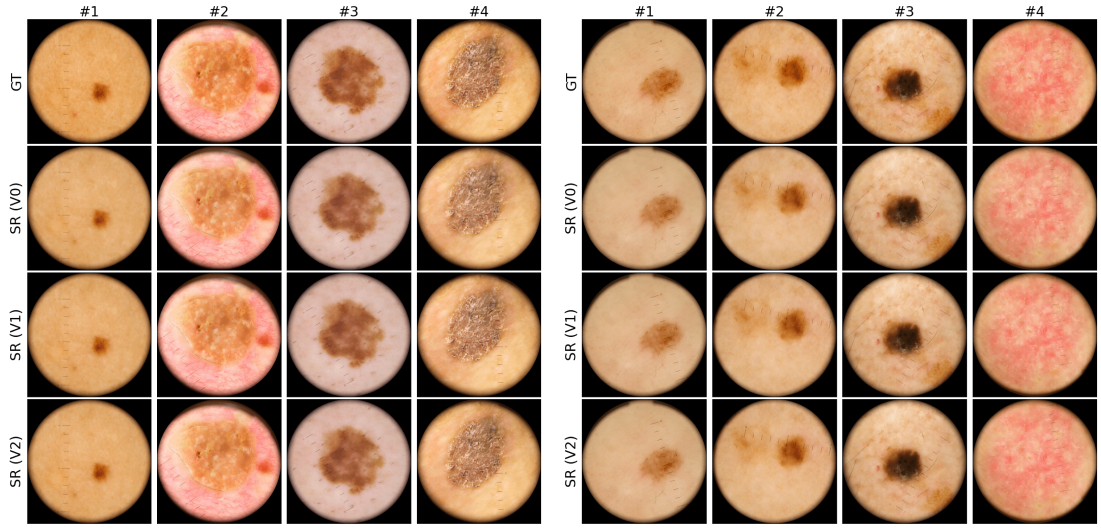
To accurately assess the performance of each model on the available test sets, image quality assessment (IQA) metrics were supported by direct visual assessments as a means of quantitatively determining the efficacy of each model respect to the original ground thruth images and eventually choosing the proper model among the three. Therefore, model gained from training version 0, 1 and 2 were confronted.



(a) ISIC **(b)** Color Constancy (CC) ISIC

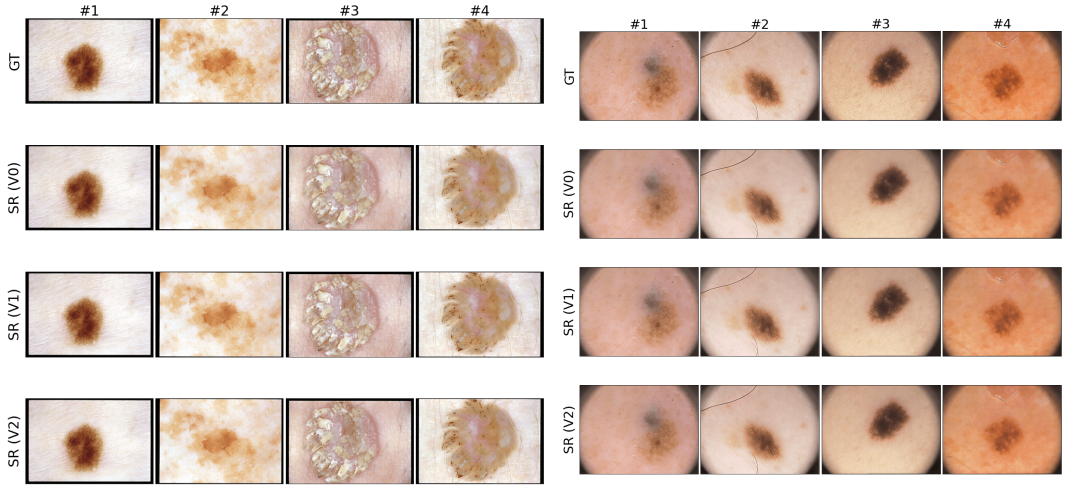
Figure 3.5: Visual Direct Assessment to evaluate the impact of Super-Resolution Image Restoration on normal and color-normalized images.

It is apparent through visual observation that all three models perform relatively well. Nonetheless, Model Version 1 exhibit superior performance compared to Version 0 and 2, as it is able to restore images from a more complex degradation space and produce images that are sharper. It is noteworthy that Model Version 1's outputs are generally smoother and clearer than the original ground truth (GT), resulting in an improvement in image quality by mitigating artifacts caused by noise, blur, and compression. Nevertheless, an overall analysis indicates that the restored images exhibit occasional deficiencies in certain color and texture details compared to the original ground truth in specific image regions, which could result in a loss of diagnostic information and potentially lead to inaccurate results in subsequent processes.



(a) Novara Good Quality (GQ) **(b)** Novara Bad Quality (BQ)

Figure 3.6: Visual Direct Assessment grid plot comparison on Good and Bad quality Novara datasets.



(a) Atlas **(b)** PH2

Figure 3.7: Visual Direct Assessment grid plot comparison on Atlas, PH2 and Nurugo Test Sets.

3.2.4 Comparison: LR, GT, SR

As a further step in the analysis, four images from the Novara Good Quality Test Test Set were selected to be displayed in a grid format, as shown in Figure 3.8, for direct visual evaluation of the differences among:

- **Low Resolution images (LR):** images that were degraded using the same degradation pipeline involved in the Training Set
- **Ground Thruth images (GT):** the original images that were downloaded and left untouched.
- **Super Resolution images got from inferencing GT images (SR from GT):** high-quality images obtained by applying the trained GAN model to restore the GT image.
- **Super Resolution images got from inferencing LR images (SR from LR):** high-quality images obtained by applying the trained GAN model to restore the LR image.

Furthermore, a comprehensive analysis utilizing IQA metrics was conducted on the entire Novara GQ Test Set to complement the visual assessment provided by the grid plots. The aim of this analysis was to gain a general understanding of:

- the effectiveness of the degradation process outlined in Chapter 2.2.4, used to generate the paired dataset, by comparing the original images (GT) and their degraded counterparts (LR).
- the dissimilarity between restoring the original, undistorted image (SR from GT) and restoring its low resolution counterpart (SR from LR). In this case, the original GT images were also involved as pristine images to understand the effectiveness of the restoration process and the extent to which the restored SR images preserve significant features while maintaining similarity to the original GT images.

The results of the comparison, both in terms of visual appearance and through the utilization of IQA metrics, reveal that the LR images demonstrate a reduction in image quality, providing evidence that the employed degradation pipeline effectively covers a broad range of real-world degradation scenarios. Additionally, the SR images restored from the original, undistorted images (GT) exhibit a higher quality, with greater texture and feature complexity, both visually and in terms of metrics, compared to the SR images restored from the low resolution (LR) images. This indicates that utilizing a high-quality starting image is likely to result in a more robust and detailed restored image, and that the performance of the model is still limited by the quality constraints of the dataset. Lastly, the SR images reconstructed

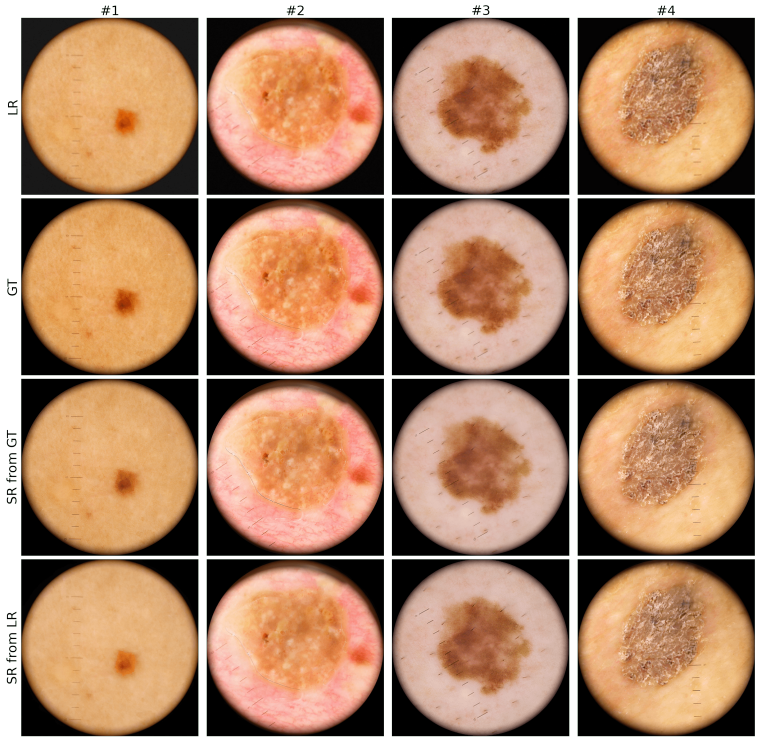
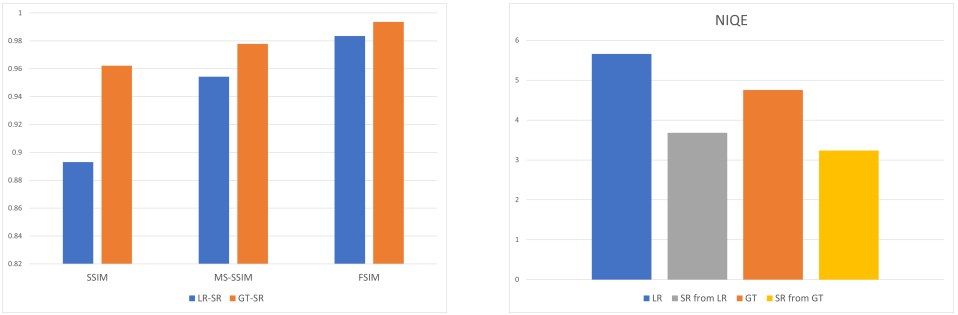


Figure 3.8: Grid plot of four images from Novara GQ dataset, showing the difference among normal and low degradated images as well as their respective restored copies.



(a) SSIM, MS-SSIM, FSIM **(b)** NIQE

Figure 3.9: Image Quality Assessment, both Full and No Reference metrics, on Novara Good Quality (GQ) Test Set to compare the difference between inferring starting from a degradated low-resolution image and a normal one (LR-SR vs. GT-SR).

from both the ground truth and low-resolution images exhibit improved quality compared to their starting quality counterparts, as determined through both metric-based and visual inspection assessments, indicating that the trained model performs adequately in the restoration task for which it was implemented.

3.2.5 Texture Analysis

The Novara GQ Test Set was also utilized to compute a variety of color and texture features with the aim of gaining a more complete understanding of the results obtained from the IQA metrics and visual assessment. This analysis focused on the area surrounding the lesion to evaluate the impact of the restoration process on the characteristics of the images related to the lesion, which is essential in ensuring an accurate diagnostic result. The analysis of the restored image samples in

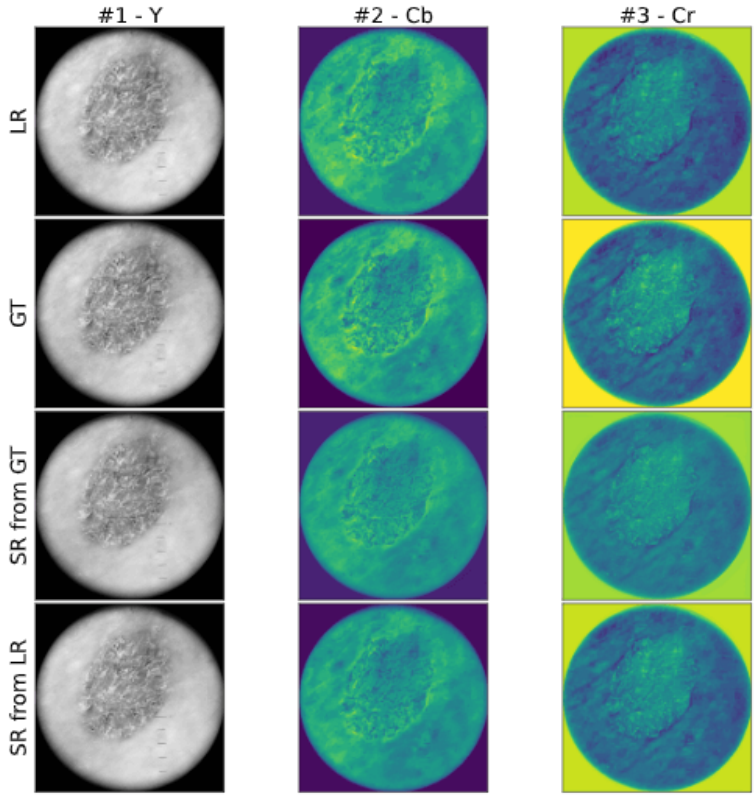


Figure 3.10: Grid plot of an image from Novara GQ dataset, showing the 3 channels belonging to the YCbCr color space: it shows both color (Y) and texture (Cb and Cr) features among the four type of processed images.

comparison to their starting input samples, as depicted in Figure 3.10, reveals

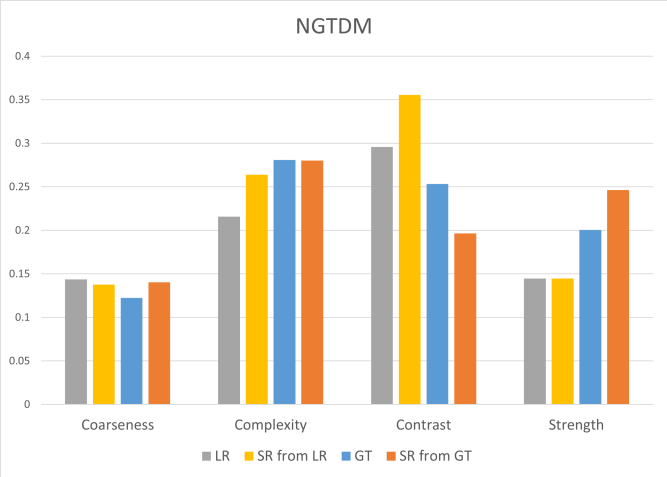


Figure 3.11: Some NGTDM features calculated from Novara GQ dataset.

that the restored images exhibit greater clarity and smoothness. Specifically, the Luminance (Y channel) exhibits an improvement in terms of smoothness, resulting in a more defined and less noisy output image. This is particularly evident when examining the Chrominance (Cb, Cr channels), which demonstrate a clearer yet valuable and significant texture information. It is important to note that while the restored images exhibit an overall improvement in clarity and smoothness, the Texture Analysis visualization reveals that there may be a loss of some color and texture features in certain areas. This could potentially be problematic in diagnostic tasks, particularly if it involves important regions of the image such as the lesion area, as this loss of information may lead to inaccurate or incomplete diagnoses. Thus, it is crucial to consider the trade-off between image quality and preservation of diagnostic information when utilizing super-resolution techniques in medical imaging. Furthermore, the degradation process can be evaluated by comparing the LR output with the GT one, and it is apparent that the Low Resolution output is of sufficient low quality to be considered as a suitable starting point to emulate real-world scenario degradations.
In support of the visual assessment, the extracted NGTDM features, graphed in Figure 3.11, suggest that the coarseness of an image is primarily affected in the GT-SR (ground truth and super-resolution) couple, with the restored image displaying a greater degree of uniformity. Additionally, low-resolution (LR) images exhibit lower coarseness and thus, a greater degree of uniformity. When evaluating the complexity of an image, it is primarily affected in the LR-SR couple, with the SR image displaying a higher degree of complexity, characterized by rapid texture changes. Conversely, the degradation process results in a reduction of complexity features. Contrast is one of the most affected features in both degradation and

restoration processes. SR images tend to exhibit greater contrast when compared to their respective LR versions. Conversely, if the starting image is the GT, the contrast tends to be lower, indicating that starting from a less degraded image leads to a more smooth and clear result. Finally, the strength of an image is relatively consistent in the case of the LR-SR couple, but tends to increase when restoration is performed on GT images. This suggests that starting with a higher quality image leads to a result with more distinct primitives and better-defined features.

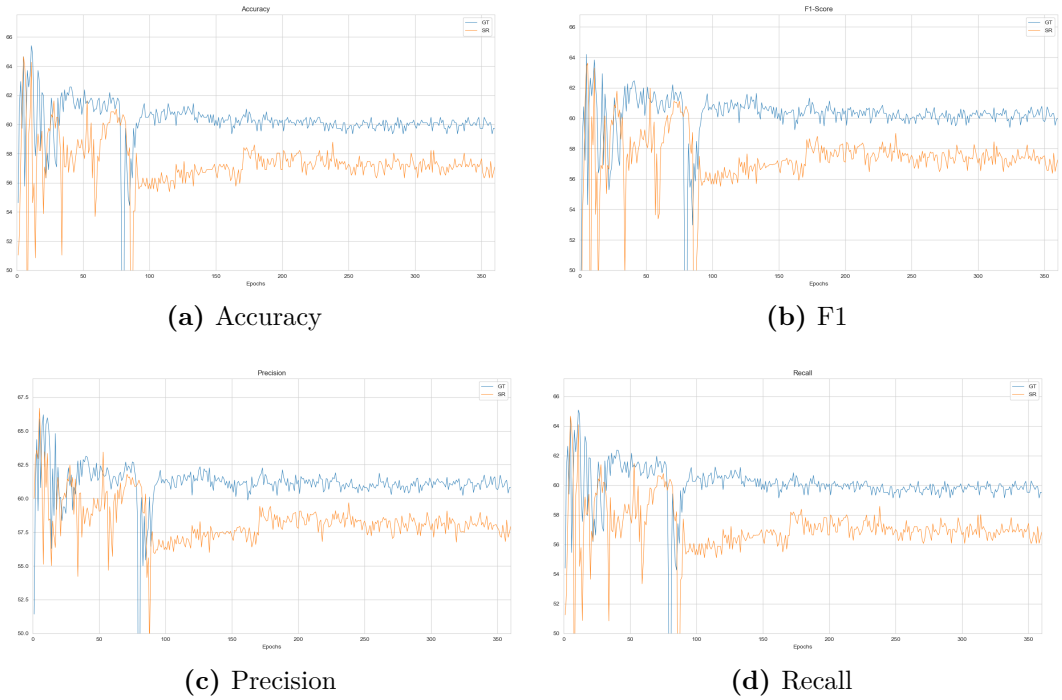
3.2.6 Skin Lesion Classification

A classification network was employed to evaluate the image restoration performance of Real-ESRGAN on the ISIC Test Set. The ISIC Test Set was selected for classification as it includes label classification data, allowing for a supervised multiclassification approach. Additionally, this Test Set does not include any images that were used in training the Real-ESRGAN. The network was trained and tested twice, using identical parameters, except for the input images, which were either the restored (SR) images or the original (GT) images. The performance of the ResNet50 through the epochs, on the training and validation sets, are illustrated in Figure 3.12.
The results depicted in Figure 3.12 demonstrate that the network does not exhibit significant improvement during the training epochs. However, the overall metrics on the test set are presented in Table 3.2, which indicate that the network trained on the ISIC SR images achieved slightly higher average metrics score on the Test Set when compared to the network trained on the ISIC GT images. Further

	Accuracy	Precision	Recall	F1-Score
Ground Truth (GT) ISIC	62.98	61.50	62.63	60.78
Super Resolution (SR) ISIC	65.34	66.08	65.33	64.28

Table 3.2: Overall classification metrics of the ResNet50 in identifying skin lesions on the Test set. It include metrics such as accuracy, precision, recall, and F1-score, which provide a comprehensive understanding of the model's performance.

experiments are necessary to establish the validity of these results. This may involve utilizing a larger and more robust dataset, as well as evaluating other architectures, such as EfficientNet-B4. These architectures, as illustrated in Figure 3.13, have demonstrated superior performance while retaining similar size complexity. Such evaluations would provide a more comprehensive and valid understanding of the true impact of the restoration process on the classification of skin lesions.



(a) Accuracy **(b)** F1

(c) Precision **(d)** Recall

Figure 3.12: Classification performance of the ResNet50 in identifying skin lesions on the Validation set, evaluated over the course of the training epochs.

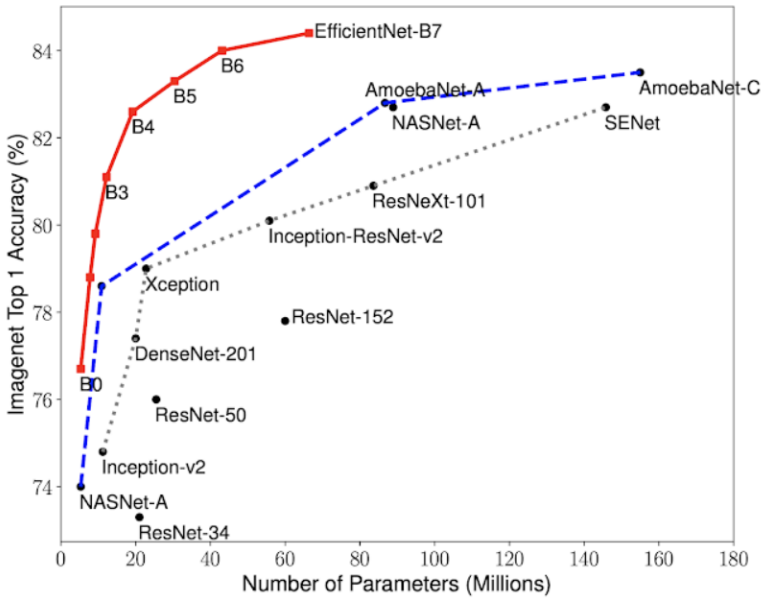


Figure 3.13: Graph depicting the accuracy performance of multiple state-of-the-art architectures as a function of the increasing size of the model [54].

Chapter 4

Conclusion

4.1 Further Work

Further work is necessary to more thoroughly assess the quality and realism of synthetic images, specifically in the context of medical imaging. This includes evaluating the performance of the implemented model by comparing it to state-of-the-art architectures and examining its ability to adapt and generalize to images from various medical domains. These studies will furnish valuable insights regarding the strengths and limitations of the model and will provide guidance for forthcoming development.

Mean Opinion Score (MOS)

Mean Opinion Score (MOS) testing is a widely used subjective image quality assessment method that involves human raters assigning perceptual quality scores to images. These scores are usually on a scale from 1 (low quality) to 5 (high quality). In the MOS testing procedure, participants are asked to select the images they perceive as realistic from a collection containing both real and generated images. The final MOS is calculated as the arithmetic mean of the ratings provided by the human raters. Despite its widespread use, MOS testing has several inherent defects, such as non-linear perception of scales, biases and variance in rating criteria, and differences in subjective views between different raters. However, when the number of evaluators and evaluations is sufficient, MOS can still be a reliable image quality assessment method.

Human Turing Test

To subjectively evaluate the realism of synthetic images, qualitative Turing experiments can be conducted. In these experiments, certified dermatologists may be

asked to choose images relevant to a given skin condition from a set containing both real ground truth images and restored images. This approach can provide additional information about the accuracy and realism of the restored images, particularly when it comes to their relevance to specific skin conditions.

Applications to other Medical Domains

It may be interesting to conduct an experiment in which the implemented and fine-tuned dermoscopic Real-ESRGAN model is applied to images from other medical domains, such as MRI and CT scans, retinal images, and microscopy images. These images may present unique characteristics that could impact the model's performance:

- **Color space:** images from different medical domains may use different color spaces, which can impact the way they are restored from the Super Resolution GAN. For example, dermoscopic images and retinal images are typically colored, while MRI and CT scans are grayscale, thus may be more challenging to process, since the GAN model has been trained on colored images.
- **Image resolution:** The resolution of images from different medical domains can vary significantly. Higher resolution images typically contain more detailed information, but may also be more challenging to process due to their larger size.
- **Image modality and content:** the content of images from different medical domains can vary significantly due to their different imaging modalities, which use different physical principles to generate images, thus affecting the appearance and characteristics of the images. For example, dermoscopic images contain detailed information about skin structures, while MRI and CT scans show internal organs and tissues. Microscopy images may show very small structures at high magnification.

Conducting this experiment could provide valuable insights into the adaptability and generalizability of the model to different types of medical images.

Comparison with State-of-the-Art Architectures

Evaluating the fined-tuned dermoscopic Real-ESRGAN against similar state-of-art architectures can be a highly interesting task for several reasons. Firstly, it allows researchers and developers to benchmark the performance of their custom GAN against established architectures and gain insights into its strengths and weaknesses. This can help identify potential areas for improvement and guide future development efforts. Secondly, comparing a custom GAN to state-of-art

architectures can provide valuable insights into the broader landscape of GAN research and development. It can help identify trends and patterns in the field, and highlight areas where further research and development are needed. Finally, evaluating a custom GAN against similar state-of-art architectures can also help to establish its credibility and demonstrate its effectiveness to a wider audience. By demonstrating that the custom GAN performs competitively or even outperforms established architectures, researchers and developers can build confidence in their approach and encourage others to consider using it in their own work.

4.2 Conclusions

This study introduces a Generative Adversarial Network (GAN)-based approach for reconstructing high-resolution skin lesion images from low-quality ones. The proposed approach employs image processing techniques and data augmentation, followed by a customized degradation pipeline to prepare the paired dataset. The Real-ESRGAN architecture is then utilized to train the model.
The performance of the resulting model is evaluated using multiple Image Quality Assessment (IQA) metrics such as FSIM and NIQE, employing a diverse range of test sets consisting of dermoscopic images that possess different characteristics from the dataset used for training. Furthermore, the synthetic restored images are compared with ground truth images from the ISIC Test Set in terms of their efficacy for skin lesion classification using a deep network.
The findings of this study suggest that the proposed GAN-based approach shows promise for reconstructing high-resolution skin lesion images from low-quality ones and further research in this area may be beneficial.

Bibliography

[1] Kim JY and Dao H. *Physiology, Integument*. Treasure Island (FL): StatPearls, May 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK554386/ (cit. on pp. 1, 3, 4).

[2] Madhero88 and M.Komorniczak. «Skin layers». In: (). URL: https://en.wikipedia.org/wiki/File:Skin_layers.png (cit. on p. 2).

[3] Martini Frederic H., Tallitsch, Robert B., Nath, Judi L., M.D. Ober, William C., R.N. Ober, and Claire E. *Human Anatomy*. 9th ed. Pearson College Div, 2017 (cit. on pp. 1, 3).

[4] Giuseppe Micali, Daniele Innocenzi, Gabriella Fabbrocini, Giuseppe Monfrecola, and Antonella Tosti e Stefano Veraldi. *Le basi della dermatologia*. 1st ed. Springer Milano. URL: https://doi.org/10.1007/978-88-470-2065-8 (cit. on pp. 3, 4).

[5] «Cancer Surveillance Branch (CSU)». In: (). URL: https://gco.iarc.fr/ (cit. on pp. 5, 11).

[6] «Skin Cancer Treatment». In: *National Institute of Health (NIH): National Cancer Institute (NCI)* (). URL: https://www.cancer.gov/types/skin/hp/skin-treatment-pdq (cit. on pp. 4–6).

[7] D. De Berker, J.M. McGregor, B.R. Hughes, and on behalf of the British Association of Dermatologists Therapy Guidelines and Audit Subcommittee. «Guidelines for the management of actinic keratoses». In: *British Journal of Dermatology* 156.2 (2007), pp. 222–230. DOI: https://doi.org/10.1111/j.1365-2133.2006.07692.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2133.2006.07692.x (cit. on p. 6).

[8] C.A. Morton, A.J. Birnie, and D.J. Eedy. «British Association of Dermatologists' guidelines for the management of squamous cell carcinoma in situ (Bowen's disease) 2014». In: *British Journal of Dermatology* 170.2 (2014), pp. 245–260. DOI: https://doi.org/10.1111/bjd.12766. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjd.12766 (cit. on p. 6).

[9] Gandhi SA and Kampp J. «Skin Cancer Epidemiology, Detection and Management». In: (2015). URL: https://doi.org/10.1016/j.mcna.2015.06.002 (cit. on p. 6).

[10] Beatrous SV, Riahi RR, Grisoli SB, and Cohen PR. «Associated conditions in patients with multiple dermatofibromas: Case reports and literature review». In: *Dermatology Online Journal* (). DOI: https://doi.org/10.5070/D3239036479 (cit. on p. 7).

[11] «Moles». In: (). URL: https://www.mayoclinic.org/diseases-conditions/moles/symptoms-causes/syc-20375200 (cit. on p. 8).

[12] Woodhouse JG and Tomecki KJ. «Common benign growths». In: (Aug. 2010). URL: https://www.clevelandclinicmeded.com/medicalpubs/diseasemanagement/dermatology/common-benign-growths/ (cit. on p. 9).

[13] «Melanoma Treatment». In: *National Institute of Health (NIH): National Cancer Institute (NCI)* (). URL: https://www.cancer.gov/types/skin/hp/melanoma-treatment-pdq (cit. on pp. 10, 12).

[14] Marco Rastrelli, Saveria Tropea, Carlo Rossi, and Mauro Alaibac. «Melanoma: Epidemiology, Risk Factors, Pathogenesis, Diagnosis and Classification». In: *In vivo (Athens, Greece)* 28 (Nov. 2014), pp. 1005–1011. (Cit. on p. 10).

[15] «Melanoma Cutaneo». In: *Fondazione Italiana per la Ricerca sul Cancro-AIRC* (). URL: https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/melanoma-cutaneo (cit. on p. 10).

[16] Qiuyu Jin. «ABCDEFG of melanoma». In: *DermNet* (). URL: https://dermnetnz.org/topics/abcdes-of-melanoma# (cit. on p. 12).

[17] «Melanoma». In: *DermNet* (). URL: https://dermnetnz.org/cme/lesions/melanoma (cit. on p. 13).

[18] Sonthalia S, Yumeen S, and Kaliyadan F. *Dermoscopy Overview and Extra-diagnostic Applications*. Treasure Island (FL): StatPearls, Jan. 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK537131/ (cit. on p. 13).

[19] «Dermatoscope Heine Delta 20T». In: (). URL: https://www.heine.com/it/prodotti/dermatoscopi-e-documentazione-digitale/dermatoscopi/dettaglio/28744-dermatoscopio-heine-delta-20t (cit. on p. 14).

[20] G Campos-do-Carmo and M Ramos-e-Silva. «Dermoscopy: basic concepts». In: *Int J Dermatol* 47.7 (2008), pp. 712–9. DOI: 10.1111/j.1365-4632.2008.03556.x (cit. on pp. 14, 15).

[21] Ashfaq A. Marghoob. «Differences between polarized and non polarized optics principles». In: (). URL: https://dermoscopedia.org/File:PD_optics_principles.jpg (cit. on p. 15).

[22] dermoscopedia: Ofer Reiter, Florentia Dimitriou, Alon Scope, Ash Marghoob, and Ralph Braun. «Differences between polarized and non polarized dermoscopy: dermoscopedia». In: (2022). URL: https://dermoscopedia.org/w/index.php?title=Differences_between_polarized_and_non_polarized_dermoscopy&oldid=17037 (cit. on p. 15).

[23] Gabriella Campos-do-Carmo and Marcia Ramos-e-Silva. «Dermoscopy: basic concepts». In: *International Journal of Dermatology* 47.7 (2008), pp. 712–719. DOI: https://doi.org/10.1111/j.1365-4632.2008.03556.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-4632.2008.03556.x (cit. on pp. 16, 17).

[24] Victor Wiley and Thomas Lucas. «Computer Vision and Image Processing: A Paper Review». In: *International Journal of Artificial Intelligence Research* (2018) (cit. on p. 18).

[25] K. R. Castleman. *Digital Image Processing* (cit. on p. 18).

[26] Shweta Kadam. «Difference beetwen Human Vision and Computer Vision». In: (). URL: https://medium.com/analytics-vidhya/cnn-series-part-1-how-do-computers-see-images-32462a0b33ca (cit. on p. 19).

[27] Christian Janiesch, Patrick Zschech, and Kai Heinrich. «Machine learning and deep learning». In: *CoRR* abs/2104.05314 (2021). arXiv: 2104.05314. URL: https://arxiv.org/abs/2104.05314 (cit. on pp. 19–21).

[28] L. Alzubaidi, J. Zhang, A.J. Humaidi, et al. «Review of deep learning: concepts, CNN architectures, challenges, applications, future directions». In: *J Big Data* 8 (2021), p. 53. DOI: 10.1186/s40537-021-00444-8 (cit. on pp. 19, 21, 23–26, 48, 59).

[29] Ravindra Parmar. «Training Deep Neural Networks». In: (). URL: https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964 (cit. on p. 20).

[30] Li Z, Koban KC, Schenck TL, Giunta RE, Li Q, and Sun Y. «Artificial Intelligence in Dermatology Image Analysis: Current Developments and Future Trends». In: (2022). DOI: 10.3390/jcm11226826. URL: https://pubmed.ncbi.nlm.nih.gov/36431301/ (cit. on pp. 22, 59).

[31] Ankan Dash, Junyi Ye, and Guiling Wang. «A review of Generative Adversarial Networks (GANs) and its applications in a wide variety of disciplines - From Medical to Remote Sensing». In: *CoRR* abs/2110.01442 (2021). arXiv: 2110.01442. URL: https://arxiv.org/abs/2110.01442 (cit. on pp. 26, 27, 32, 33).

[32] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. «Super-resolution image reconstruction: a technical overview». In: *IEEE Signal Processing Magazine* 20.3 (2003), pp. 21–36. DOI: 10.1109/MSP.2003.1203207 (cit. on p. 28).

[33] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. *Deep Learning for Image Super-resolution: A Survey*. 2019. DOI: 10.48550/ARXIV.1902.06068. URL: https://arxiv.org/abs/1902.06068 (cit. on pp. 28–30, 59).

[34] «Bicubic Interpolation». In: (). URL: https://en.wikipedia.org/wiki/Bicubic_interpolation (cit. on pp. 29, 30).

[35] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. «Image Super-Resolution Using Deep Convolutional Networks». In: *CoRR* abs/1501.00092 (2015). arXiv: 1501.00092. URL: http://arxiv.org/abs/1501.00092 (cit. on p. 31).

[36] Christian Ledig et al. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*. 2016. DOI: 10.48550/ARXIV.1609.04802. URL: https://arxiv.org/abs/1609.04802 (cit. on pp. 31, 32).

[37] «Latent diffusion architecture used by Stable Diffusion». In: (). URL: https://github.com/CompVis/latent-diffusion (cit. on p. 34).

[38] «ISIC: The International Skin Imaging Collaboration». In: (). URL: https://www.isic-archive.com/ (cit. on p. 36).

[39] Massimo Salvi, Francesco Branciforti, Federica Veronese, Elisa Zavattaro, Vanessa Tarantino, Paola Savoia, and Kristen M. Meiburger. «DermoCC-GAN: A new approach for standardizing dermatological images using generative adversarial networks». In: *Computer Methods and Programs in Biomedicine* 225 (2022), p. 107040. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.107040. URL: https://www.sciencedirect.com/science/article/pii/S0169260722004229 (cit. on p. 36).

[40] Giuseppe Argenziano et al. «Interactive atlas of dermoscopy». In: (). URL: https://espace.library.uq.edu.au/view/UQ:229410 (cit. on p. 37).

[41] «PH2». In: (). URL: https://www.fc.up.pt/addi/project.html (cit. on p. 37).

[42] «Albumentation». In: (). URL: https://albumentations.ai/docs/ (cit. on p. 40).

[43] JIM TOL. «Image augmentation: how to overcome small radiology datasets». In: (2019). URL: https://www.quantib.com/blog/image-augmentation-how-to-overcome-small-radiology-datasets (cit. on p. 42).

[44] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. *Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data*. 2021. DOI: 10.48550/ARXIV.2107.10833. URL: https://arxiv.org/abs/2107.10833 (cit. on pp. 43–45).

[45] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. «ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks». In: *CoRR* abs/1809.00219 (2018). arXiv: 1809.00219. URL: http://arxiv.org/abs/1809.00219 (cit. on p. 45).

[46] *Imagenet*. URL: https://www.image-net.org/index.php (cit. on pp. 49, 59).

[47] Hamed Alqahtani, Manolya Kavakli, and Dr. Gulshan Kumar Ahuja. «An Analysis Of Evaluation metrics of GANs». In: July 2019 (cit. on pp. 52, 53).

[48] Anish Mittal, Rajiv Soundararajan, and Alan Bovik. «Making a "Completely Blind" Image Quality Analyzer». In: *Signal Processing Letters, IEEE* 20 (Mar. 2013), pp. 209–212. DOI: 10.1109/LSP.2012.2227726 (cit. on pp. 53, 57).

[49] Umme Sara, Morium Akter, and Mohammad Uddin. «Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study». In: *Journal of Computer and Communications* 07 (Jan. 2019), pp. 8–18. DOI: 10.4236/jcc.2019.73002 (cit. on p. 54).

[50] MathWorks. *Image Quality Metrics*. URL: https://it.mathworks.com/help/images/image-quality-metrics.html (cit. on p. 57).

[51] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. «Blind image quality evaluation using perception based features». In: *2015 Twenty First National Conference on Communications (NCC)*. 2015, pp. 1–6. DOI: 10.1109/NCC.2015.7084843 (cit. on p. 58).

[52] A. Drimbarean and P.F. Whelan. «Experiments in colour texture analysis». In: *Pattern Recognition Letters* 22.10 (2001), pp. 1161–1167. ISSN: 0167-8655. DOI: https://doi.org/10.1016/S0167-8655(01)00058-7. URL: https://www.sciencedirect.com/science/article/pii/S0167865501000587 (cit. on p. 58).

[53] Pyradiomics. *Radiomic Features*. URL: https://pyradiomics.readthedocs.io/en/latest/features.html# (cit. on p. 58).

[54] Staff Software Engineer Mingxing Tan and Google AI Quoc V. Le Principal Scientist. *EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling*. URL: https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html (cit. on p. 73).