Università degli Studi di Salerno

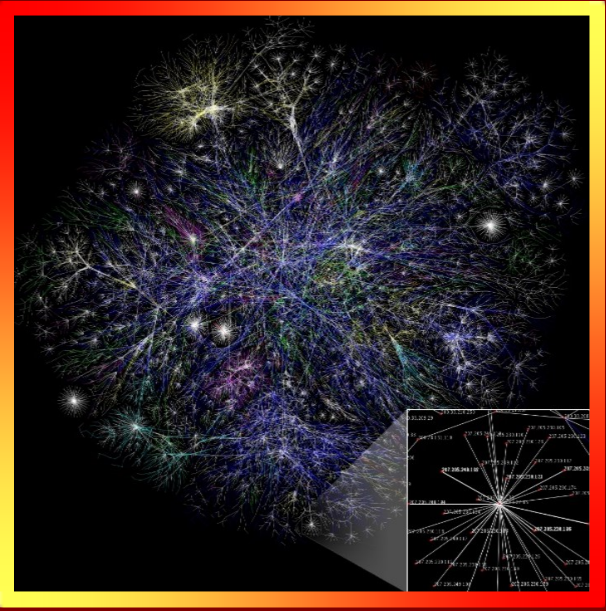
Dipartimento di Informatica

Dottorato di Ricerca in Informatica – XXXV Ciclo

Tesi di Dottorato/Ph.D. Thesis

The use of Artificial Intelligence for the Melanoma Binary Classification Problem on images (MIBCP)

Luigi Di Biasi



Supervisor: **Prof. Genoveffa Tortora**

Ph.D. Program Director: **Prof. Andrea De Lucia**

AA 2021/2022

Curriculum Computer Science and Information Technology

Università degli Studi di Salerno

Dipartimento di Informatica

Dottorato di Ricerca in Informatica

Curriculum Computer Science and Information Technology

XXXV Ciclo

Tesi di Dottorato / Ph.D. Thesis

The use of Artificial Intelligence for the Melanoma Binary Classification Problem on images (MIBCP)

Luigi DI BIASI

Supervisor: **Prof. Genoveffa TORTORA**

PhD Program Director: **Prof. Andrea DE LUCIA**

A.A 2021/2022

Serena means serenity, clear, cloudless.
Dedicated to the my natural anxiolytic

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Genoveffa Tortora, for her invaluable guidance, support, and encouragement throughout my research. Her expertise and insight have been instrumental in shaping my work. Also, I would like to express my heartfelt gratitude to the memory of Professor Michele Risi, who made a profound impact on my academic and personal development.

I am also deeply grateful to my colleagues Fabiola De Marco and Alessia Auriemma Citarella for their helpful discussions, feedback, and assistance during this project. Their contributions have been essential to the success of my research.

Furthermore, I extend special thanks to Stefano Piotto, Lucia Sessa and to the entire laboratory for their invaluable support and the knowledge they imparted to me during my time there. The expertise I gained from them has been tremendously helpful in my doctoral studies. I would like to dedicate this work to my parents Aurelio and Assunta and to my siblings, who have made significant sacrifices to support me throughout my academic journey. Their unwavering support has been a constant source of motivation and inspiration.

Finally, I want to extend my thanks to all the individuals who provided me with support and encouragement along the way. In particular, I want to express my gratitude to my wife Serena Lembo and children Aurelio e Francesco, who have shown incredible patience, understanding, and support during this challenging period. Their love and unwavering encouragement have been a constant source of strength and inspiration.

ABSTRACT

The healthcare industry plays a critical role in saving lives every day. As a result, researchers, physicians, and experts are constantly working to find new ways to address illnesses and disabilities. In addition, technological advancements, especially in artificial intelligence and machine learning, have helped the scientific community design and propose advanced diagnostic tools to help physicians make crucial patient care decisions. These tools allow researchers to analyze vast amounts of data in new ways, often in real-time, for various purposes, such as detect-ing patterns behind illnesses, analyzing signals and detecting potential cancer from images.

In this context, this work was dedicated to the study of the Melanoma Image Binary Classification Problem (MIBCP), mainly by analyzing and proposing solutions to addressing the open issues in this field that did not allow a massive utilization of computer-aided diagnostic systems for early diagnosis. In particular, this work focuses on the resolution of the problems that may be behind high-performance automatic prediction models: the need to minimize risk situations, even by accepting lower overall performance; the opportunity to use clinical images instead of instrumental images in early diagnosis; the need able doctors to evaluate how the automatic prediction models learn and choose the results, rather than blindly relying only on the statistical values that can be calculated by analyzing the performance of the system on training, validation and test tests; the need for a scalable architecture specialized in allowing the refinement of prediction models in a fast and accessible way to non-experts.

The results reported aim to help increase trust in the automatic system that can be implemented thanks to deep learning, in particular by showing these systems' advantages, limitations and disadvantages and providing tools that show the potential to overcome these limitations. Also, this work aims to improve Melanoma early detection, which is now a limiting factor for first-line therapies in this tumour pathology.

ABSTRACT IN ITALIANO

Il settore sanitario svolge un ruolo fondamentale nel salvare vite ogni giorno. Di conseguenza, ricercatori, medici ed esperti lavorano costantemente per trovare nuovi modi per affrontare malattie e disabilità. Inoltre, i progressi tecnologici, in particolare nell'intelligenza artificiale e nell'apprendimento automatico, hanno aiutato la comunità scientifica a progettare e proporre strumenti diagnostici avanzati per aiutare i medici a prendere decisioni cruciali sulla cura del paziente. Questi strumenti consentono ai ricercatori di analizzare grandi quantità di dati in modi nuovi, spesso in tempo reale, per vari scopi, come rilevare modelli dietro malattie, analizzare segnali e rilevare potenziali tumori dalle immagini. In questo contesto, questo lavoro è stato dedicato allo studio del Melanoma Image Binary Classification Problem (MIBCP), principalmente analizzando e proponendo soluzioni per affrontare le questioni aperte in questo campo che non hanno consentito un utilizzo massiccio di sistemi diagnostici assistiti da computer per la diagnosi precoce . In particolare, questo lavoro si concentra sulla risoluzione dei problemi che possono essere alla base di modelli di previsione automatica ad alte prestazioni: la necessità di minimizzare le situazioni di rischio, anche accettando prestazioni complessive inferiori; l'opportunità di utilizzare immagini cliniche invece di immagini strumentali nella diagnosi precoce; la necessità di medici in grado di valutare come i modelli di previsione automatica apprendono e scelgono i risultati, piuttosto che affidarsi ciecamente solo ai valori statistici che possono essere calcolati analizzando le prestazioni del sistema su traning, validation e test set; la necessità di un'architettura scalabile specializzata nel consentire l'affinamento dei modelli di previsione in modo rapido e accessibile ai non esperti. I risultati riportati mirano ad aumentare la fiducia nel sistema automatico che può essere implementato grazie al deep learning, in particolare mostrando vantaggi, limiti e svantaggi di questi sistemi e fornendo strumenti che mostrano il potenziale per superare questi limiti. Inoltre, questo lavoro mira a migliorare la diagno-

si precoce del melanoma, che ora è un fattore limitante per le terapie di prima linea in questa patologia tumorale.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The main objective of this introduction is to provide a brief overview of the context around the melanoma detection problem (Section 1.1). Then the importance of using Artificial Intelligence (AI) techniques to construct increasingly efficient Computer-Aided Design (CAD) systems is described. Also, some critical open challenges in this field are presented in Section 1.2. Finally, the contribution and the organization of the thesis are described (Section 1.3 and Section 1.5).

1.1 overview

Healthcare saves lives every day. Nowdays, thousand of people, among researchers, physicians and experts are trying to discover new ways to address illness and disabilities. The technological progress of recent years, particularly in the Artificial Intelligence (AI) and Machine Learning (ML)/Deep Learning (DL) fields, concurrently with the emerging trend of Open Data and the explosion of Cloud Architecture, has helped the scientific community to discover new potential ways to address issues in the healthcare area, allowing a drastic reduction in the time needed to perform experimentations. Technology can revolutionize how we view medicine and is vital to developing advanced diagnos-tic tools which enable crucial patient care decisions [62]. In the past, using artificial intelligence by researchers and physicians was bound to many limitations. Finding quality data, informa-tion, and specialized technology to perform adequate prediction model training was particularly difficult. As a result, for many years, performing artificial intelligence research and application was possible only thanks to big projects or research groups of BigTech companies.

Fortunately, we are still experiencing a favourable time regarding the quality of the research tools and data we can access. In particular, we have witnessed the success of Big Data concurrently with the explosion of new parallel computation paradigms, like CUDA and OpenCL, while we already possess robust distributed computation technologies like MPI and OpenMP. Together, these tools allow researchers to analyze the enormous amount of data provided by BigData in new ways, often in real-time, in particular with the rise of Cloud Computing: looking for a hidden pattern behind illness, analyzing in real-time electroencephalography (EEG) or electrocardiogram (ECG) looking for hidden malfunctions, detecting potential cancer from images; the number of new applications is very high.

Developing high-quality, accurate Artificial Intelligence resources showed the potential to improve the work of the clinicians by intervening in the prevention, diagnosis and treatment of many pathologies. Some modern AI and standard computer science technologies have increased clinical laboratory power of devices, allowing many diagnostic activities outside laboratories. In addition, medical aids and new advanced diagnostic equipment are increasingly relying on qualified experts in the field to supplement medical evaluations and assist in diagnosis. Scientific progress in information and computer technology has led to a new trend in Healthcare: Digital Health and Telemedicine.

While healthcare saves lives, digital healthcare shows the potential to save a tremendous amount of life. Nowadays, we have the potential to change the conception of medicine dramatically, and, at the same time, we are starting to design and test advanced diagnostics systems in making decisions intrinsic to patient care. The hope is to have in the nearest future CAD that can improve early diagnosis.

In this context of growing in automatic tool proposal for early diagnosis, this work focused on Melanoma disease, particularly on the Image Binary Classification Problem (MIBCP), understood as the possibility chose from a skin lesion image an answer between benign and malignant.

This study explicitly addresses the challenges when high-performing automatic prediction models might be introduced in real healthcare scenarios:

- minimizing potential risks related to automatic choices, even lowering the overall performance;

- exploring the possibility of using clinical images for early diagnosis instead of relying solely on instrumental images (dermoscopic, histologic);
- allowing the doctor to assess (and understand) how the prediction models work;
- make decisions to increase confidence in CAD, instead of forcing them to trust on statistical values derived from performance testing.

In addition, the study highlights the requirement for a scalable architecture that is user-friendly and accessible for non-experts, allowing for effortless refinement of prediction models.

Melanoma is a type of aggressive skin cancer developing from melanocytes, cells that produce melanin pigment, and is one of the deadliest tumours in the world. Skin cancer is defined as the unregulated expansion of skin cells caused by DNA damage. In the early stages, it might be confused with a normal naevus. Although it represents a minority of cutaneous malignancies, this tumour is the leading cause of mortality. Therefore, the importance of an early diagnosis of Melanoma has become increasingly evident, especially in subjects with a high risk of developing cancer because it allows us an increment of cure rate.

Furthermore, early detection is critical for first-line therapy in this tumour pathology. Generally, in clinical practice, a first visual inspection by a dermatologist is used to diagnose Melanoma, often with polarized light magnification's assistance. However, the attestation of a correct diagnosis also depends on the ability of physician related to his/her degree of experience in discriminating between skin lesions.

Nowadays, many proposals exist for CAD systems for dermatologists. However, the claims that incredible performance of the AI surpasses performance of clinicians still have not resulted in massive use of CAD systems for early diagnosis: this is because many other challenges exist, for example, the little trust in these systems, the potential risks related to automatic choices, the need for specific hardware, the need for computer science skills.

Despite the high accuracy reached by CAD, the final word on a cancer diagnosis is still delegated to the Biopsy: early diagnosis remains challenging due to the need for histological analysis to ensure correctness in diagnosis; nowadays, patients still have to undergo surgery to minimize false positives (FP) and negatives (FN) in diagnosis; this fact contrasts with all the potential the technologies provide to us. Also, most of this software involves computer vision(CV) related techniques like border detection, symmetry/asymmetry analysis, colour analysis and dimension. Furthermore, it relies on dermoscopic images, and then they require specific hardware to acquire the images: the need for specific hardware often slows down the speed in broad utilization of these techniques. Also, as discussed in the next chapter, CAD performance can drop while the underlying training set changes, which may drastically impact the diagnosis.

In order to address these issues, this work relies on the assumption that, while designing a CAD system that will work in the health field, it might be mandatory to focus on reaching the lowest false negative rate (FNR) possible to avoid errors that can lead to life-threatening situations. Also, the work relies on the assumption that the CAD system must enable the users to understand what a decision is based on, mainly using explainable AI (XIA) concepts, and must allow not skilled users to train and test new prediction models. Finally, the clinical images are considered the primary source of information for this project due to their potential large availability and simplicity in acquisitions.

1.2 open challenges

The solution to a clinical issue is closely linked to medical research and the expertise of healthcare professionals. Therefore, the best outcome can be achieved only through collaboration between technological and medical actors.

Many CADs have been developed and proposed to aid dermatologists in determining if a skin lesion is or could become a melanoma, but there are still challenges to be addressed. First, increasing trust in this system in the final user (doctor, phisicians, dermatologist) is mandatory.

Most CAD utilizes computer vision techniques, such as border detection, colour analysis, and dimension detection, to analyze skin lesions. However, most of this software uses images, and the complexity of skin images can pose problems such as irregular fuzzy boundaries, noise, low contrast, or poor lighting, which must be considered in developing melanoma detection systems. So then, the complexity of the images is still an open issue in this research field.

Building CAD systems can be broken down into main several steps:

1. a dataset of melanoma and non-melanoma images is obtained or created. The images may be dermoscopic, clinical, or histological, with the dermoscopic images providing the most detail but with a smaller dataset, while clinical images are less detailed but more readily available, and histological images provide the highest image resolution;

2. one or more computer vision and image processing techniques extract features from the images, serving as the training inputs;

3. the classification model can be trained using neural networks, SVMs, custom predictors, classifiers, or other machine/deep learning techniques, and multiple techniques may be combined to improve accuracy;

4. finally, validation and test steps are performed to measure the performance of the model.

Another open issue is selecting the best suitable classification model that can provide better performance regarding melanoma classification. Also, it is essential to note that the term "performance" definition must not be only related to the classification's accuracy and speed. Focusing on the False Positive and False Negative is essential because a wrong diagnosis may lead to a life-threatening situation.

In addition, other open issues exist when dealing with CAD and Melanoma: the first is the storage space and computational power required to train complex models on large datasets to perform well. The second issue is the effort needed to update one or more models. Furthermore, once a model is trained and deployed, there is no simple way to improve its performance without retraining the model: these issues are clearly explained by what happened in the ISIC 2018 challenges; the performance of the winning model from the ISIC 2018 challenge dropped from 88.5% to 63.6% in the ISIC 2019 challenge due to the introduction of more categories and images; this highlights the fact that deep learning performances are strictly related to the quality of the image and the structure of the training dataset.

Therefore, managing inter-class and intra-class dissimilarities is an open issue that can profoundly impact the performance of the classification system that must be addressed.

1.3 contribution of this thesis

The first contribution of this work is an overview of the melanoma classification problem analyzed from two different points of view (POV): the physician and the data scientist POVs. In particular, the first chapter provides an overview of melanoma disease and the current clinical assessment standards. From the physician and clinical POV, having the lowest false negative rate (FNR) is more important than having the highest accuracy because the false negatives may lead to a life-threatening situation in patients. It is essential to note that the physician's POV (PPOV) is used to drive all the design and evaluation of the proposed models. Also, an overview of the machine learning and deep learning approaches proposed to build a reliable CAD system is provided.

The second contribution is related to the opportunity to provide a classification model that can use clinical images instead of dermatoscopic and histological images. Nowadays, we can process and analyze medical images through complex mathematical algorithms to extract information and create new knowledge of pathological and physiological phenomena that are not detected with visual analysis alone. Furthermore, thanks to classification models able to use clinical images, we could potentially speed up the broad utilization of CAD systems by avoiding the need for specific hardware like dermatoscopic because mobile devices surround us with ultra-high resolution cameras. In particular, the second contribution is a deep study of the main available convolutional neural network (CNN) architectures performances from the perspective of minimizing the FNR using clinical images.

The third contribution is the proposal of new CNN models designed by genetic algorithms (GA): in particular, the GA fitness function was oriented to drive the evolution of the GA populations to minimize the CNN structures and the FNR.

The final contribution of this work is to provide a robust system against performance drops that can occur while the training data changes: in particular, a cloud-fog-edge architecture is proposed, and the experimentation results are presented in the latest chapters.

This thesis aims to help to improve Melanoma early detection, which is now a limiting factor for first-line therapies in this tumour pathology.

1.4 research questions

In the context of melanoma detection from digital images, it is more important to consider the false-negative rate (FNR) than global accuracy because of the severe consequences of missing a melanoma diagnosis. Melanoma is a deadly form of skin cancer that can spread quickly, making early detection crucial for successful treatment. If a melanoma is missed by a detection system, this could lead to a delay in diagnosis and treatment, which can have devastating consequences for the patient's health outcomes.

While global accuracy is a useful metric for measuring the overall performance of a detection system, it may not be sufficient for assessing the effectiveness of a system for detecting rare or critical cases such as melanoma. A high global accuracy score may suggest that the system is effective, but if the FNR is high, it means that a significant proportion of melanomas are being missed. Therefore, minimizing the FNR is a critical objective in melanoma detection to ensure that patients receive timely and accurate diagnoses.

The research question for this doctoral thesis is centered around the development of a reliable and accurate melanoma detection model using clinical images. Specifically, the study aims to investigate the performance of various convolutional neural network (CNN) architectures in minimizing the false negative rate (FNR) for the detection of melanoma. The study will focus on the use of clinical images rather than dermatoscopic and histological images, as this has the potential to broaden the utilization of computer-aided diagnosis (CAD) systems by eliminating the need for specialized hardware.

The research will involve a comprehensive evaluation of the performance of different CNN architectures, with a focus on achieving optimal FNR while also considering the overall accuracy, precision, and recall of the model. Also, the research investigated how GA can be used to generate new CNN architectures optimized to minimize FNR.

The results of this study are expected to contribute to the development of more effective and accessible melanoma detection tools, which could have significant implications for the early diagnosis and treatment of this deadly disease.

The research questions can be summarized as follows:

> What is the performance of convolutional neural network (CNN) architectures for minimizing the false negative rate (FNR) when using clinical images instead of dermatoscopic and histological images? How can the use of clinical images potentially speed up the utilization of computer-aided diagnosis (CAD) systems? What is the performance of the proposed convolutional neural network (CNN) models designed by genetic algorithms (GA) in minimizing both the structure of the CNN and the false negative rate (FNR), and how do these models compare to existing CNN architectures in terms of accuracy, precision, and recall? How effective is the proposed cloud-fog-edge architecture in providing a robust system against performance drops caused by changes in training data, and what are the experimental results of its implementation?

1.5 organization of the thesis

After this Chapter Introduction, the thesis is divided as follows:

In Chapter *2*, there is an overview of skin cancers, Melanoma and detection issues. In particular, it reports an overview of the scientific literature relating to melanoma detection problems using clinical images (MDCI). The Chapter begins with a review of some processes required for melanoma detection and ends with a description of the classification algorithms proposed in the literature to address these issues.

Chapter *3* illustrated the experimental results related to multiple CNN architectures trained on clinical images with and without segmentation and data augmentation in order to obtain the best model of CNN and for the minimization of False Negative Rate (FNR).

Chapter *4* presents an alternative way to use an extended version of GA to address the MDCI. In particular, the Chapter presents the experimental results obtained using GA (selection, mutation, merging and crossover) to perform the design of a CNN driven by the GA scoring function; the maximization of the prediction accuracy and the minimization of the FNR was used as scoring functions.

In Chapter *5*, the contribution of the intra-class dissimilarities (ICD) and extra-class similarity (ECS) presence in melanoma images dataset in affect classification performance is reported; then, a hybrid architecture design on the continuous re-training approach is presented and analyzed.

Finally, conclusions and future studies are followed in Chapter *6*. This Chapter highlights the contributions proposed by these works and any future directions.

2 BACKGROUND

This Chapter describes the critical point regarding MDCI. In particular, some notions on melanoma and its assessment are outlined, considering advantages and disadvantages (Section 2.1, Section 2.2). Then the most diffused methods of machine learning and deep learning for melanoma detection are described (Section 2.3 and Section 2.4). In particular, we focus on machine and deep learning techniques and genetic algorithms and evaluation performance (Section 2.5).

2.1 melanoma

Melanoma is a form of skin cancer that starts from melanocytes, the cells that produce melanin pigment [82]. It belongs to the generic skin cancer class, including basal cell carcinoma, squamous cell carcinoma and Melanoma [5]. From the biological point of view, skin cancer is an uncontrollable proliferation of skin cells driven by DNA damage that can occur due to multiple reasons: ultraviolet radiation, sun expositions, sunburn, radiation and genotoxic effect [31]. In addition, evidence in the literature suggests that genetic and environmental factors might be considered risk factors: race, in particular caucasian, light-coloured skin and a positive family history of Melanoma can be considered risk factors. [77].

Melanoma is one of the deadliest tumours in the world and is aggressive [18]. Although it represents a minority of cutaneous malignancies, this tumour is the leading cause of mortality related to skin cancers due to its ability to metastasize to different tissues rapidly. In addition, the site of the primary Melanoma's initial growth strongly correlates with the risk of metastatic progression, with melanomas developing in the head, neck, and trunk having a higher risk of metastatic progression than those developing in the limb [82].

Skin melanoma cases have significantly grown during the past 30 years. Nevertheless, the trends vary depending on the age group: For instance, between 2007 and 2016, the rate for those under 50 decreased by 1.2% annually, while the rate for people over 50 increased by 2.2%. The American Cancer Society estimates that alone in the United States, there will be 100350 new cases and 6850 deaths in both sexes in 2020 [1].

Like other tumours, a five-stage protocol is used to assess melanoma development. The protocol is based on the American Joint Committee on Cancer Staging Manual [30]: the primary characteristics used to distinguish between benign or malignant lesions are thickness, ulceration, and metastasis to lymph nodes or other regions of the body; for tumour stage between I and III, the excision is the treatment of choice. The Figure 2.1 shows the stages of melanoma in detail.

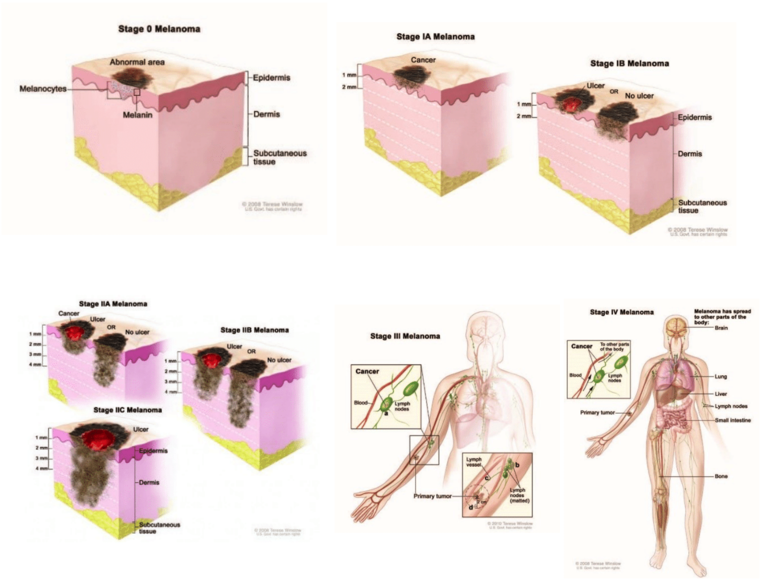
In stage 0, the epidermis contains abnormal melanocytes that could become cancerous. Therefore, in this Stage, the Melanoma is considered in situ.

In the next Stage, thickness and ulceration might be considered. In particular, Stage I is divided into Stage IA and Stage IB: in the first one (IA), ulcerations are not present, and the thickness of the tumour is less than 1 mm; in the second one (IB), ulcerations may be present for the tumours with thickness less o equal than 1mm; Also, it is considered to be into Stage IB tumours without ulcerations but with thickness is between 1mm and 2mm.

Stage II is split into three substages: IIA, IIB and IIC; In the first one (IIA), the tumour could have ulceration, but the thickness must be less or equal to 1mm, or the tumour think can be between 1 and 2mm without ulcerations; In the second one (IIB), the tumours are ulcerated, and the thickness is between 2mm and 4mm; the tumour with the thickness greater than 4mm that not exposes ulceration is considered in Stage IIB; If the tumour thickness is greater than 4mm without ulceration, it is considered in Stage IIC.

The tumour is considered Stage III if cancer has spread to one or more lymph nodes without progressing to other body regions:

[1] https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf



Credits: Terese Winslow & Us Gov.

Figure 2.1: Stages of Melanoma.

in this Stage, thickness and ulceration are not considered variables of primary importance. Starting from Stage III, the process called metastasis starts, and excision of the lesion alone is not sufficient to ensure healing: In this case, chemotherapy [18], radiation therapy [93], immunotherapy [81] and targeted therapy [10] are required treatments.

If the tumour has progressed to other body regions, it is considered in Stage IV. The following Table 2.1 reports a summary of the melanoma stages.

Due to the death ratio of Melanoma, it is critical to establish the tumour stages quickly and with certainty to evaluate the best therapy to improve the prognosis.

2.2 melanoma assessment and visual inspection

Usually, a first visual inspection of a dermatologist is used to discriminate between benign skin lesions and potential Melanoma

•

Thickness	Ulceration	Metastasis	Other Body Region	STAGE
NA	NA	False	False	0
<= 1mm	False	False	False	IA
<= 1mm	True	False	False	IB/IIA
]1mm, 2mm]	False	False	False	IB/IIA
]2mm, 4mm]	True	False	False	IIB
> 4mm	False	False	False	IIB
> 4mm	True	False	False	IIC
NA	NA	True	False	III
NA	NA	True	True	IV

Table 2.1: Summary of the melanoma stages

situations. Often, the visual inspection is made with polarized light magnification's assistance, with a dermatoscopic. The attestation of a valid diagnosis given by visual inspection is entirely dependent on the ability of the pyisician in discerning between distinct skin lesions, which is proportional to his/her level of expertise. The Biopsy, on the other hand, has the last say on the cancer diagnosis.

The strict interconnection between the ability of the dermatologist and the diagnosis may lead to life-threatening situations.

For example, after a visual inspection, a physician may confuse a Stage IB Melanoma with Stage IIA melanoma: these stages may be confused due to overlapping (see Table 2.1) in thickness range and ulceration presence or missing; the real problem in this scenario is that the next Stage after IB is the Stage IIB, the last Stage before metastasis starts. Therefore, failing to determine the correct Tumour Stage with the visual inspection may reduce cure success. In particular, considering the Stage system defined by American Joint Committee on Cancer Staging Manual, visual inspection may be enough only for the Melanoma stage between 0 and IIC: when the phase of metastasis and aggression of other sites has not yet started. This fact strongly shows why the research must focus on providing systems and techniques to improve early diagnosis as best as possible.

Another dangerous situation may arise if a physician incurs false positive (FP) or false negative (FN) diagnoses: in the first situation, the patient may be doubly stressed due to the fear of Melanoma and the need to face with a biopsy, but in the end, he will realize that he has no cancer in his body. However, in the FN situation, the patient is already ill and may have Melanoma in the initial Stage that be cured entirely; unfortunately, in the FN scenario, the physician diagnosed that the skin lesion is benign, and he does not require any further analysis.

Apart from the stress, in the FP scenario, the patient will survive. However, in the FN scenario, the evolution of Melanoma may lead to an infaust prognosis. This fact suggests that research efforts should be focused on providing diagnosis protocols that minimize FN events at the expense of many more FPs.

A dermatologist performing a Visual Skin Inspection (VSI) may face several difficulties, mainly because every skin lesion could significantly differ from or resemble the others, and it is common **to have extra-class similarities (ECS) and intra-class differences (ICD)**: the ECS and ICD are crucial in the experimentation which is the basis of the contributions of this thesis.

The presence of areas with anatomical-morphological traits strikingly similar to those of a benign nevus can confuse skilled dermatologists: this fact makes the photo screening a difficult task, and there is a big chance that errors will be made throughout the inspection.

For example, Figure 2.2 reports multiple examples of different Melanomas and naevus that may confuse during preliminary visual Inspection.

In order to overcome confusion situation, a dermatologist might use different techniques. A dermatologist can evaluate a skin lesion following different inspection protocols:

- **ABCD rule**, which considers a lesion’s Asymmetry, Border, Color and Diameter [64];
- **The Seven Point Checklist (7PCL)**, which considers the change in the size of the lesion, irregular pigmentation, irregular border, inflammation, itch or altered sensation, larger than other lesions (diameter >7mm), and Oozing/crusting of the lesion [38];

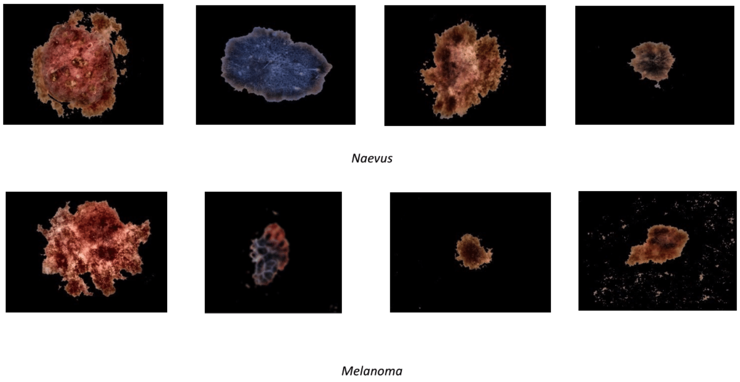


Figure 2.2: Several examples of neavus and melanoma images.

- **The Weighted 7PCL**, which considers the same features as 7PCL but assigns more weight to the size of the lesion, the irregularity in the pigmentation and border are considered more important.

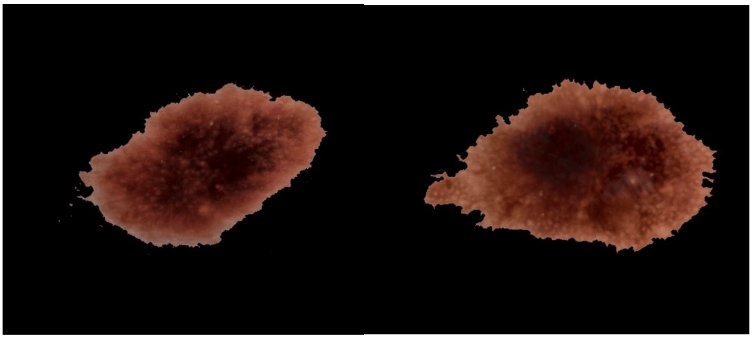
However, the dermatologist may only speculate a diagnosis: the final word on cancer must be confirmed by a biopsy.

2.2.1 Dermoscopic and Histological Inspection

As described above, the most superficial inspection can be performed by observing the skin without the support of any tool. For example, the dermatologist might try to apply one of the inspection protocols described above for each suspect lesion visible on the body of the patient.

Unfortunately, ICD and ECS may confuse a dermatologist, particularly during the feature analysis following one of the available inspection protocols. For example, Figure 2.3 shows a melanoma and a simple naevus that show similar colour, thickness and diameter. Also, it is essential to consider that noise is often present in images, and there may not always be enough contrast between tissues to allow for a precise diagnosis.

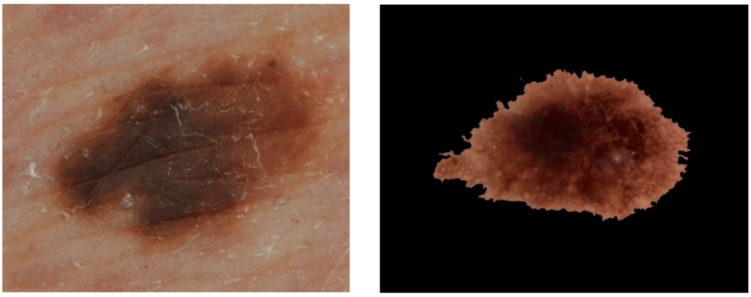
In order to improve the quality of the analysis, a dermatologist may use a dermoscopy during the VSI: dermoscopic images



Benign *Malignant*

Figure 2.3: Example of simple naevus with similar colour, thickness and diameter.

have a higher informative content that can be extracted using the "ABCD" rule. Also, dermoscopic images allow us the examination of the pigment arrangement in the context of the lesion, the depth of localization, and the presence of subcutaneous patterns visible to the human eye via simple clinical analysis. Finally, dermoscopic images consent the retrieval of lesion enlargement between 10 and 20 times. The Figure 2.4 shows the differences between a clinical image (a portable device image), and a dermatoscope image.



Clinic *Dermoscopic*

Figure 2.4: Example of clinical and dermoscopic image.

The highest informative content can be extracted from the histological images: histological images result from suitably treated biological tissue preparation for subsequent microscopic analysis. The histological examination allows us to have information on the morphology and functionality of the tissues and cells that constitute them. As a result, histological images provide the highest informative content at the cellular level.

In order to avoid relying only on physician skills, it is possible to use many approaches.

2.2.2 Lymph Node Mapping

Since the flow of lymph is directed, some cancers spread predictably from where the cancer started: this orderly progression of some cancers' spread usually begins with regional lymph nodes and then moves on to the next echelon of lymph nodes. For example, the most typical location for malignant melanoma metastases is lymph node metastasis [11].

In some cancer situations, if cancer spreads, it will eventually affect the lymph nodes (lymph glands) nearby the tumour before moving on to other body regions. In that case, there is a substantial possibility that cancer has not spread if the sentinel lymph node is cancer-free: sentinel lymph nodes are lymph nodes that may have metastasized, and locating one is crucial since it indicates how far the tumour has spread.

The **sentinel Lymph Node Mapping**(LNM) may be performed to ascertain whether the malignancy has spread or not: the procedure is performed using dyes and radioactive materials. Figure 2.5 report an example of a sentinel lymph node related to breast cancer: the axillary lymph nodes may be the first lymph nodes to be impacted in breast cancer since they get 75% of the lymph from the breasts.

The method does have some disadvantages, mainly when applied to melanoma patients. Only patients with positive nodes can benefit therapeutically from this method[manca ref].

A false negative result could come from failing to find cancer cells in the sentinel node since there might still be malignant cells in the lymph node basin. Furthermore, there is no convincing proof that patients who undergo a complete lymph node

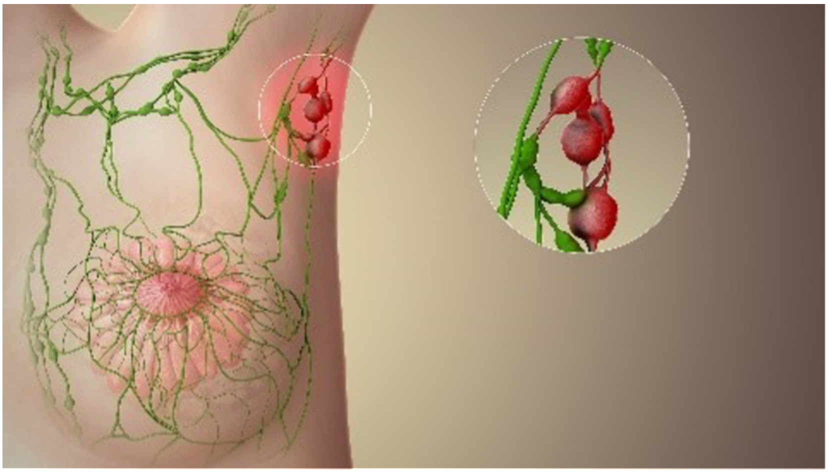


Figure 2.5: Sentinel lymph node.

dissection in response to a positive sentinel lymph node result have an improved prognosis in comparison to patients who wait until later in their disease, when the lymph nodes can be felt by a doctor, to undergo a complete dissection. These patients may undergo unnecessary complete dissection, increasing their chance of developing lymphedema [91].

2.2.3 Computer Tomography

Although the most typical location for malignant melanoma metastases is lymph node metastasis, lung metastasis was found in 85% of end-stage melanoma patients, making the lungs and pleura the second most prevalent site for malignant melanoma metastases after lymph node involvement. Approximately 60–70% of melanoma patients had liver metastases at the time of autopsy, making it the most prevalent metastasis involving the abdomen and pelvis. In that case, a computerized x-ray imaging procedure known as a **Computed Tomography** (CT) scan can be used instead of LNM.

CT uses a narrow x-ray beam that is quickly spun around the patient's body to produce signals processed by the machine's computer to produce "slices", known as tomographic images. In order to create a three-dimensional (3D) image in which we may identify tumours or other abnormalities in the patient's body, these slices can be digitally "stacked" together after being collected. A CT scanner employs a motorized x-ray source that spins around the circular opening of a doughnut-shaped frame called a gantry: this is in contrast to a traditional x-ray, which uses a stationary x-ray tube.

During the CT procedure, the patient is lying on a bed that rotates across the gantry as a narrow beam of x-rays is shot into the body by the x-ray tube: particular digital x-ray detectors are placed immediately across from the x-ray source and are used in CT scanners in place of film. The detectors catch the x-rays as they leave the patient and send them to a computer [11].

The CT computer constructs a two-dimensional imaging slice of the patient using advanced mathematical methods each time the x-ray source completes one full revolution. Depending on the CT equipment being used, the thickness of the tissue shown in each imaging slice might change, but it typically ranges from 1 to 10 millimetres. After finishing a whole slice, the image is saved, and the motorized bed is slowly lowered onto the gantry. Then another image slice is created by repeating the x-ray scanning procedure. This procedure is repeated until the required number of slices has been gathered.

The computer can either display the image slices separately or stack them to create a 3D image of the patient that displays the skeleton, organs, tissues and any anomalies the doctor hopes to spot. This approach has various benefits, including the ability to rotate the 3D image in space or to see slices one after the other, which makes it simpler to pinpoint the precise location of a potential problem.

Figure 2.6 show an example of CT images reporting Melanoma liver metastasis.

2.2.4 Positron Emission Tomography

Malignant tumours metabolize glucose faster than benign tumours. So then, it might be possible to discriminate between benign and malignant tumours by observing the metabolic activity of the cells [27].

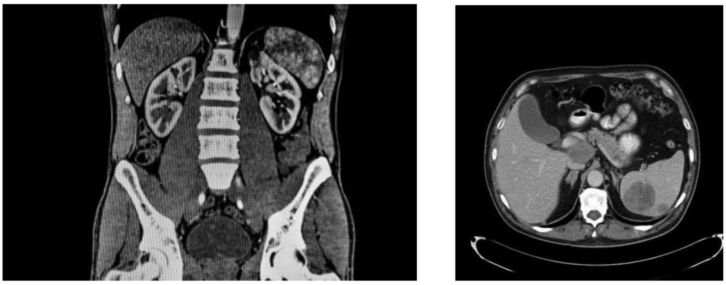


Figure 2.6: Melanoma metastasi in the liver.

The metabolic activity can be measured using a process known as **Positron Emission Tomography (PET)** [63] which belongs to nuclear medicine. PET can assess physiological function by examining neurotransmitters, blood flow, metabolism, and radiolabeled medicines [9]: also, it provides quantitative studies that make it possible to track relative changes over time as a disease process alters or reacts to a particular stimulus.

In the melanoma scenario, PET can gauge how quickly the body uses glucose in various locations by using radiolabelled glucose analogue 18-fluorodeoxyglucose (FDG) that can be accumulated to estimate the rate of glucose consumption. Often, whole-body scans are used to stage and detect melanoma metas-tases [39].

PET is crucial for portraying the biomedical changes to the body, the structure and function of the organs or tissues, and their biochemical characteristics. As a result, PET allows detection of the onset of a disease process before other imaging procedures can detect structural alterations such as CT.

A radiopharmaceutical—also known as a radionuclide or radioactive tracer—is used in the PET to assess the metabolism of a particular organ or tissue. For example, the method finds radioactivity after injecting a tiny amount of a radioactive tracer into a peripheral vein: the tracer, which is often labelled with oxygen-15, fluorine-18, carbon-11, or nitrogen-13. The overall radioactive dose is comparable to the computed tomography dose.

A PET scan lasts 10 to 40 minutes. The patient is fully clothed during computed tomography and painless. An example is shown in the Figure 2.7.

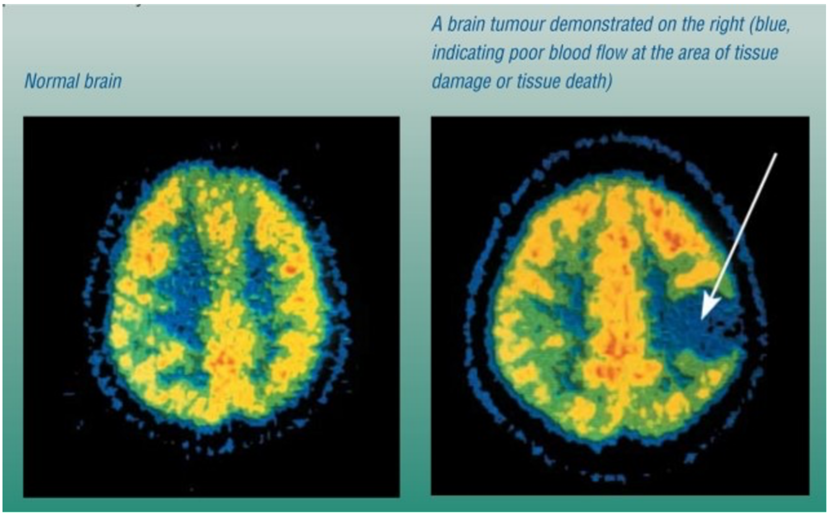

Normal brain
A brain tumour demonstrated on the right (blue, indicating poor blood flow at the area of tissue damage or tissue death)

Figure 2.7: Melanoma PET scan.

2.2.5 Magnetic Resonance Image

Melanoma cells may expose well-defined patterns if exposed to a magnetic resonance field. For example, in [59] twenty-seven sites of Melanoma were evaluated with **Magnetic Resonance Image** (MRI) technique discovering four signal patterns that can be used to determine if a melanoma is present or not. Also, for some malignant Melanoma, in particular, intracerebral malignant Melanoma, PET can be used with MRI to increase sensitivity in detection [97]. Figure 2.8 shows an example of melanoma detection through MRI.

MRI uses radiofrequency radiation that transitions between the nuclear spin states of tissue hydrogen atoms (protons) caused by a high external magnetic field. The variance in relaxation durations among various tissue types is the primary source of contrast in clinical MRI. Therefore, the spin-lattice (Tl) and spin-spin (T2) relaxation times are frequently used to describe an MR signal's behaviour and can be used as a baseline for generating

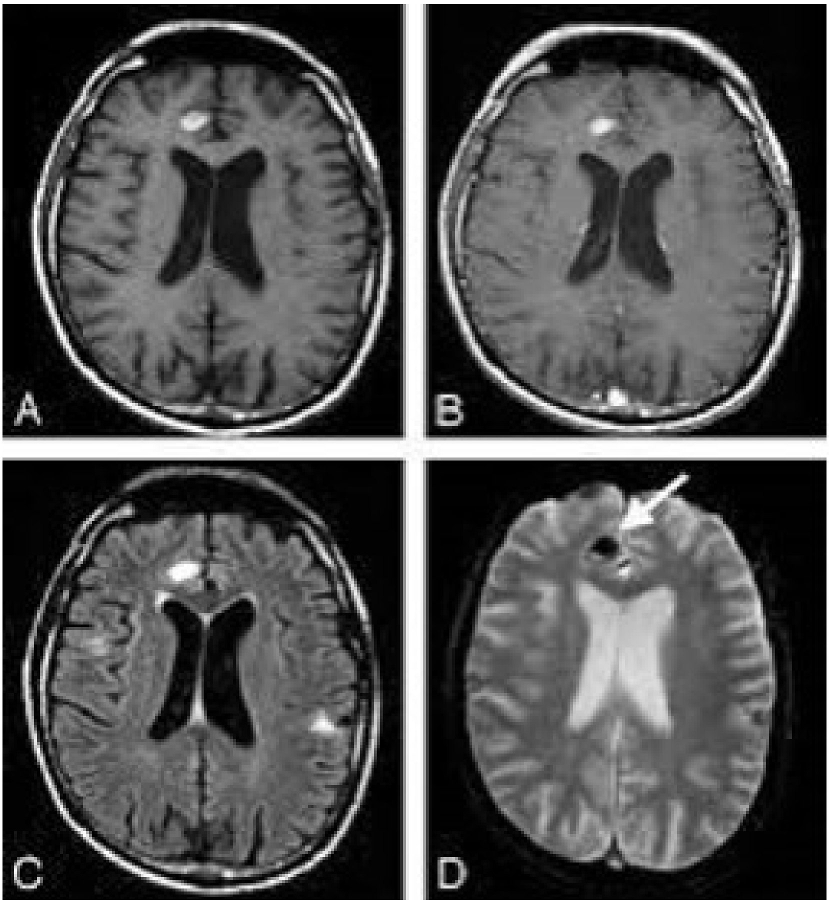


Figure 2.8: Melanoma metastases detect by MRI [29].

MR contrast. Therefore, it is possible to use fat and muscle signal intensities as reference tissues to grade lesions into high, low or intermediate intensity categories.

Consequently, MRI offers further data for locating and describing tumours and healthy tissues.

2.2.6 Standard blood chemistry tests

Standard blood chemistry testscan be used to look for a marker. A few biomarkers are produced by melanoma cells in the blood to identify lymph node metastasis. The most important genes associated with Melanoma are MLANA2 (Melan-A; also known as MART-1, melanoma antigen recognized by T cells, MAGEA3 (melanoma antigen family A, 3), and MITF (microphthalmia-associated transcription factor) [52].

There are many serological prognostic markers for cutaneous Melanoma, such as S100 calcium-binding protein B (S100B), melanoma inhibitory activity (MIA), hepatocyte growth factor (HGF), eosinophil cationic protein (ECP), serum indoleamine 2,3-dioxygenase (IDO), decreased vitamin D level, and serum lactate dehydrogenase (LDH). Beyond these, we also have DNA methylation biomarkers, microRNA (miRNA), Long noncoding RNAs (lncRNAs) and Histone modification biomarkers, as summarized in Figure 2.9.

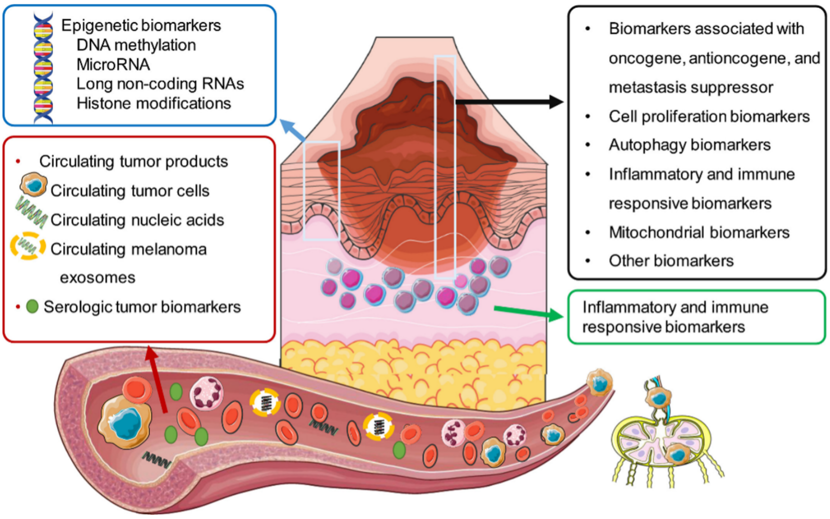

Epigenetic biomarkers
DNA methylation
MicroRNA
Long non-coding RNAs
Histone modifications
• Circulating tumor products
Circulating tumor cells
Circulating nucleic acids
Circulating melanoma exosomes
• Serologic tumor biomarkers
• Biomarkers associated with oncogene, antioncogene, and metastasis suppressor
• Cell proliferation biomarkers
• Autophagy biomarkers
• Inflammatory and immune responsive biomarkers
• Mitochondrial biomarkers
• Other biomarkers
Inflammatory and immune responsive biomarkers

Figure 2.9: A summary of the sierological prognostic markers for cutaneous Melanoma.

Circulating tumour cells (CTCs) have been identified using various methods, including filtration, flow cytometry, microfluidics, and the microbead sorting technique, but the most common method is RT-PCR of the Melanoma associated transcripts. Circulating biomarkers with liquid biopsies from melanoma patients may be used for diagnosis, staging, and prognosis. The main problem is that histological evaluation is still being used because these biomarkers are only slowly validated [48, 68].

2.3 state of the art and related works

One of the most popular strategies for utilizing obtained biological data and facilitating diagnosis is CAD that support clinical decisions. Adopting a computerized system that supports the dermatologist and has strong repeatability and stability can lead to a quicker diagnosis and higher accuracy. In order to extract relevant data for digital health, these cutting-edge technologies are increasingly applied to biomedical challenges [60, 83], such as proteomics [6, 15], genetics and image and signal data classification [19, 20], and visualization [14]. Additionally, the Internet of Medical Things (IoMT), a subset of the Internet of Things (IoT) dedicated to the connectivity of all medical equipment, expands as more medical devices are connected [92]. As a result, new intelligence systems for health and well-being supported by mobile apps, robots, and remote servers such as in [16, 25] are possible. Nowadays, many proposals exist for CAD to help with melanoma detection: most of these systems work with images and lie in computer vision-related techniques like border detection, symmetry/asymmetry analysis, colour analysis and dimension. Moreover, these systems can use different image inputs, particularly images that can be dermoscopic, clinical or histological images. The final goal of these systems is to detect Melanoma using lesion photographs as input automatically. The ability of CNNs are crucial [42] in the detection [65], segmentation, and categorization of melanocytic lesions and they have been well documented [103].

However, it is possible to perform melanoma detection without images, but only by using the numerical description of the features: as an example, a CAD might be implemented by using a decision tree that uses lesion size and colour. It is possible to use decision trees [104], Support Vector Machines (SVM) [32], logistic regression [90] and Bayesian classifiers [80]; it is possible to consider this case as the ML approach. Of course, it is possible to use ML and DL separately or in combination. Also, it is possible to attempt to merge ML and DL with other computer science techniques like genetic algorithms (GA) or swarm intelligence algorithms.

2.4 machine learning and deep learning

ML is a research field aiming to design and implement algorithms which imitate human intelligence skills in problem-solving. As humans, we interact and learn from the environment and can generalize the results archived for a problem to solve another similar problem that has not yet been observed. Machine learning algorithms follow a similar behaviour: they observe examples (training data) and try to generalize a solution by computing generic rules that solve a particular problem; they try multiple times to optimize the generic rules following a score function; often, the score function is computed (or is related to) considering the number of correct and wrong answer produced by the algorithms. DL might be considered an evolution of ML. In ML, the model works rely on mathematical objects for which scientists expect to extract the "rules" that allow the model to generalize a problem. On the contrary, in the DL approach, we "believe" (have faith) that the neural network (NN) may learn how to generalize a problem by optimizing its layer weights; often, in DL, multiple "hidden layers" are accepted as is, as a black-box.

2.4.1 Underfitting and Overfitting

During model training, scientists tend to find suitable models by using first the training set (for the training step) and then the test set. If the accuracy is adequate, we may be tempted to raise the accuracy of the prediction by increasing or decreasing the selection features or by using feature engineering in our machine learning model. However, occasionally, our model might produce poor results: the reason for our model's low performance could be that it is either too simple or too complex to express the problem accurately. In DL/ML fields, the terms used to refer to these issues are overfitting and underfitting. In the underfitting case, the line (a generic model) does not cover all the points shown in the graph. Therefore, such a model tends to cause underfitting of data. Hence, it is also called High Bias. In the overfitting case, the predicted line covers all the points in the graph: the model memorized the training set entirely (including bias and outlier); in this situation, it could be believed that the model is excellent and comprehensive. Of course, it is not precisely accurate, though, because the projected line in the graph includes all outlier and noisy points. Due to its intricacy, such a model is also accountable for predicting poor results. Hence, it also goes by the name High Variance. The middle graph shows a pretty good predicted line. It covers most of the points in the graph and maintains the balance between bias and variance. In order to reach this balance between bias and variance, both ML and DL approaches must deliberately sacrifice the ability to memorize the entire training set to increase the ability to generalize a solution. It is no coincidence that scientists want to imitate human behaviour. Humans learn from the environment and generalize the results archived to solve similar problems that have not yet been observed. Almost always, we do not memorize all the steps to solve a specific problem, but we generalize and abstract the problem to use the learned skills in future. The quality of the training data [37] significantly impacts the model performance. The training data is generally derived from a collection of characteristics extracted for the particular problem to solve: as an example, in the MDCI problem addressed in this document:

- the "Collection" (training set) is a set of clinical images showing a skin lesion; the "training set" contains both "melanoma" and "benign skin lesion - naevus"; each object in the training set is labelled as "melanoma" or "naevus";
- the "characteristics" are the shape, the size, the colour, the diameter and all the other "image descriptors" may be computed from a single image (percentage of blue).

2.4.2 Training methodologies

Pattern recognition, computer vision, spacecraft engineering, finance, entertainment, computational biology, and many other fields use machine learning and deep learning techniques. It is possible to consider three main learning paradigms to train an ML/DL model: Supervised Learning, Unsupervised Learning and Transfer Learning. Furthermore, these approaches may be combined.

In Supervised Learning, the training set includes well-labelled examples. For each example, the model, during the training session, knows a priori the relationship between input and expected output: the model can utilize the training set to learn and infer the relationship between inputs and outputs. In this case, the model is trained using a proper technique, which modifies the model's internal status (often the status is represented by weights) and other parameters based on the data to reduce the prediction error. The reduction of the prediction error allows the model to have an acceptable ability to generalize and understand the relationship between input and output data. If the model is trained well, it can also generate predictions on unknowable facts. The most widely used supervised learning techniques are decision trees, Naive Bayes, Support Vector Machines (SVM), and classifiers based on neural network architectures.

In the MDCI problem, the training set is composed of melanoma and naevus images; for each, the model will know the correct label to assign.

In Unsupervised Learning, the model's internal state is changed to attempt to group the incoming data and group them in appropriate clusters. These algorithms use topological or probabilistic methods to learn a few features from the data. This approach is mainly used for clustering and feature reduction. The most famous unsupervised Learning are Principal Component Analysis (PCA), K-Means Clustering and Self-Organizing Maps (SOM).

A pre-trained model developed for one task is used as the foundation for a model for another task in Transfer Learning (TL): often, pre-trained models are utilized as the foundation for computer vision and natural language processing (NLP). As an example, a model trained on an extensive dataset to detect soccer players may be re-used to detect rugby player after a small re-training step on a small dataset containing examples of the last one. This method can solve problem "B" with a generic model M, which is strictly related to problem "A" for which model M was built. Most TL methodologies align the source and target domain input spaces under the premise that the domain distributions are the same [96].

In general, performance may degrade when TL is employed due to the intra-class difference between an object in problem "A" and "B". In order to improve performance, it is possible to carry out an additional step known as fine-tuning: by performing fine-tuning, the pre-trained weights of a model "A" are used as the basis instead of using random values for the training session of model "B". For example, if model"B" is more complex than model "A", the pre-trained weights can be used together with random values to perform a train only for the model part not available in model "A". In general, using TL, one of the following strategies is performed: to correct the source's marginal distribution difference, correct the conditional distribution difference, or correct both the marginal and conditional distribution differences. The mixed approach draws its basis from supervised, non-supervised Learning and TL. Semi-supervised models aim to use a small amount of training data labelled together with a large amount of input data without a label: This often occurs in real situations where data labelling is costly, and we can obtain a constant flow of data.

In this thesis, we used different approaches in order to classify the melanoma images. Specifically, our approach consider the use of some NNs and GAs. Moreover, we have also employed hybrid architectures in the field of melanoma detection. In this section, we will discuss the related works considered as a starting point for our research.

The vast majority of the works in the literature are based on the classification of dermoscopic images.

Already in 2013, Razmjooy et al. [78] proposed a new algorithm for hair removal applying canny edge detection and new features based on asymmetry and irregular border quantification to improve SVM. They reached an accuracy of 95%. In 2014, Ramezani et al. [76] used SVM employed with a threshold-based method for segmentation after applying noise removal techniques to the images. They used 187 features indicating asymmetry, border irregularity, color variation, dimension, and texture, which are reduced by applying PCA. The authors obtained an ACC of 82.2%, a specificity of 86.93%, and a sensitivity of 77%.

From 2015 to recent years, several neural network architectures, in particular CNNs have been used for the classification of melanoma. CNNs assistance improved the dermatologist's accuracy in interpreting skin cancers and may increase the acceptance of this new procedure further. A recent systematic review explores 19 studies comparing classifications between CNN-based classifiers for melanoma, which show superior or equivalent performance to clinicians, regardless of the type of input data [37].

Nasr-Esfahani et al. proposed using a deep learning system on a computer server equipped with a graphics processing unit (GPU) to detect melanoma lesions using clinical images. Clinical input images, which may involve illumination and noise effects, are pre-processed and then submitted to a pre-trained CNN that distinguishes between melanoma and benign cases in the proposed system. The collection consists of 170 non-dermoscopic images (70 melanoma, 100 naevi) from the University Medical Center Groningen's Department of Dermatology's digital image library (UMCG). The proposed system has reached 81% of accuracy [65].

Yu et al. [99] constructed a fully convolutional residual network (FCRN) for segmentation,incorporating a multi-scale contextual information integration scheme. Then they used very deep residual networks for classification on ISBI 2016 Skin Lesion Analysis Towards Melanoma Detection Challenge dataset. The reached accuracy was 85.5% while the specificity, sensitivity and AUC was 93.1%, 54,7% and 85.5% with segmentation.

Kawahara et al. proposed a multitask deep CNN trained on multimodal data using clinical, dermoscopic images and patient metadata. Using several multitask loss functions, each of which takes into account different combinations of input modalities and a seven-point checklist, their neural network generated multimodal feature vectors for image retrieval and detection of clinical discriminant regions [51].

Aldwgeri et al. combined several CNN architectures in order to improve the performance of melanoma classification on the dataset created for the ISIC 2018 challenge, known as HAM10000 (Human against Machine). This dataset contains 10015 images divided in different pathologies: Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AKIEC), Benign keratosis (BKL), Dermatofibroma (DF) and Vascular lesion (VASC). They implemented five CNNs: VGG-Net, ResNet50, InceptionV3, Xception and DenseNet121. An ensemble model with balanced multi-class accuracy of 80.1% and mean average 0.89 ROC AUC achieved the best performance.

The authors in [28] used a GoogLeNet DCNN model architecture trained on a dataset of clinical images of malignant melanoma (MM), squamous cell carcinoma (SCC), bowen disease, actinic keratosis, basal cell carcinoma (BCC), naevus cell naevus (NCN), blue naevus, congenital melanocytic naevus, spitz naevus, sebaceous naevus, poroma, seborrhoeic keratosis, naevus spilus and lentigo simplex. In particular, there are 540 malignant melanoma images, reaching an accuracy of 72.6%.

Zhang et al. have constructed an attention residual learning CNN, called ARL-CNN, to avoid the problem of little data available, extra-class similarity and intra-class variation. They based their network on an attention mechanism capable of increasing the possibility of discriminating the information available by focusing on its semantic meaning. The authors do not introduce new extra learnable layers in the network. Still, they delegate the possibility of grasping the semantic meaning to the more abstract feature maps of the higher layers. The authors use the dataset ISIC 2017 with 1320 additionally dermoscopy images, including 466 melanoma. ARL-CNN network consists of 50 layers and has obtained an ACC of 85%, a specificity of 89.6% and a sensitivity of 65.8% [100].

The study presented in [46] used a dataset of more than 12,000 skin images of malignant and benign tumors, from which they extracted 5,846 clinical images of pigmented skin lesions from 3,551 patients. The dataset contains 1,611 malignant melanoma images. This study used a Faster Region-based CNN (FRCNN) model because it consistently demonstrated good classification accuracy, robustness, and speed. The authors evaluate the classification of FRCNN into six classes: malignant melanoma and basal cell carcinoma (malignant classes), naevus, seborrheic keratosis, senile lentigo, and hematoma/hemangioma (benign classes). They achieve an accuracy of 86.2%. The accuracy, sensitivity and specificity for two-class classification (benign or malignant) were 91.5 %, 83.3 % and 94.5%, respectively.

Alizadeh et al. [3] suggest a method to classify skin cancer on dermoscopy images based on four steps: pre-processing, CNN classification, classification based on feature extraction, and final classification using the ensemble method. They have tested their approach on ISIC datasets.

Abbes et al. [2] proposed a fuzzy decision ontology-based CAD system consisting of two major steps. The first step introduces a framework for concept modeling based on extracted relevant features. The ontology is constructed with the concepts involved in the ABCD rule: asymmetry, border, color, and differential structures. As a result, the Bag-of-Words, which model these concepts, are generated using extracted features from skin lesion images. The second step classifies the lesion images, based on fuzzy decision rules with K-Nearest Neighbors approach, by indicating the risk level of an existing melanoma. They reached a sensitivity of (96%) and an accuracy of (92%) on two public datasets of 206 skin lesion images.

Ba et al. [7] proposed a multi-class CNN trained and validated using a dataset of 25,773 clinical images approved by the Chinese PLA General Hospital & Medical School's Institutional Review Board. It covers ten types of skin cancer: basal cell carcinoma (BCC), squamous cell carcinoma (SCC), including keratoacanthoma, melanoma (MM), Bowen disease (BD), actinic keratosis (AK), melanocytic naevus (MN), seborrhoeic keratosis (SK), haemangioma, including pyogenic granuloma, cherry haemangioma, sinusoidal haemangioma and angiokeratoma, dermatofibroma (DF) and wart. CNN used in [7] achieved an overall accuracy of 78.45%, and CNN-assisted dermatologists achieved greater accuracy (76.60% versus 62.78%) than non-assisted dermatologists in interpreting clinical images.

Kaur et al. [50] have developed a new deep convolutional neural network (DCNN) model for classifying skin lesions by connecting many blocks to allow ample feature information to pass straight through the network. They have named this architecture Lesion Classification Network (LCNet). In order to extract low and high-level feature information from lesions, each block of the network uses distinct parameters such as the number of kernels, filter size, and stride. Furthermore, since ISIC datasets are unbalanced, the authors use data augmentation and oversampling methods.

In our previous work [22], we evaluated three neural architectures on the MED-NODE dataset: AlexNet, GoogleNet and Google InceptionV3. In this previous work, we addressed the issue of TL and the development of a more adaptable system design that can accommodate changes in training datasets. Our findings suggest that AlexNet is the most robust network in terms of TL, without data augmentation, with mean accuracies of 78% and 89% with and without Otsu segmentation, respectively [21].

The Darwinian concept that the most suitable environment elements have a better chance of surviving and transferring their features to their progeny is followed by GAs in emulating the modes of evolution. There is a population of individuals (called chromosomes) that evolve from generation to generation using natural evolutionary processes. In evolution, there are three fundamental mechanisms:

- selection: the selection indicates the process of selection of the most promising solutions capable of generating individuals who survive in the environment;
- cross over: it is a genetic recombination operator which introduces variation in the population;
- mutation: it shifts the space of solutions, resulting in the development of new information and the recovery of knowledge that has been lost through time in the population.

The evolutionary algorithms use heuristic exploration to find novel solutions to problems where there is no complete knowledge of the search area. They then start with the original solution and alter, integrate, and evolve it until they have a better outcome.

The application of genetic algorithms has grown in popularity in NN optimization. Genetic algorithms can optimize several processes using the notion of biological evolution. The approach iterates through three stages: selection, crossover, and mutation, starting with a random population of network architectures [53]. GAs are used to improve CNN hyperparameters. For example, the number of neurons in each layer and the size and number of filters in each layer can affect the accuracy of a neural network model.

Recently, Pérez and Ventura have proposed a CNN architecture designed by a genetic algorithm that finds optimal members of an ensemble learning model. Their work suggests how genetic algorithms can find efficient architectures in the diagnosis of melanoma with performances 11% and 13% better than CNN models used in the literature [70].

With the advent of digital images, the objective is to use Computer Vision, Machine Learning, and Deep Learning to extract information from them and produce new knowledge: this enables the use of images for early diagnosis and subsequent treatment of various diseases. However, using real-world data to build melanoma risk models entails overcoming difficulties with data preprocessing, effective representation, and computing efficiency. In addition, computational difficulties exist while translating the data and training a machine learning algorithm, even when reliable patient data can be gathered for a predictive model. In order to use melanoma images in machine-learning models, images must first be retrieved and translated into a suitable format.

The size of the dataset and the sophistication of the machine learning algorithm may therefore provide computational difficulties.

Users can launch machines of various sizes equipped with prebuilt libraries for machine learning algorithms thanks to the cloud, a computer infrastructure that can be accessed via the internet. The most accurate model may be created by using this technology to assess a variety of algorithms: continuous training-test iterations may be needed to provide robust prediction models. Predictive accuracy is not the only element to consider when selecting classifiers and machine learning approaches when working with massive data or data that cannot be processed through conventional architectures; computational complexity and cost must also be considered.

In [79] Aaron et al. present a case study demonstrating the efficacy of a cloud-based method for learning from de-identified electronic health record data for melanoma risk prediction. In order to combine distributed and non-distributed computing in the cloud, the authors employed distributed processing with Apache Spark for data preprocessing and labelling and non-distributed processing with sci-kit-learn for machine learning model training: in particular, they explored logistic regression (LR), random forest (RF), and XGBoost (XGB) models to evaluate performance across the original and sampled datasets. XGBoost is an implementation of regularized gradient-boosted trees. gradient-boosted classifier achieved the best predictive performance with cross-validation (AUC = 79%, Sensitivity = 75%, Specificity = 68%). Compared to a model built on the original data, two orders of magnitude smaller datasets could achieve statistically similar or better performance with less than 1% of the training time and cost.

In [47], the authors used HAM10000 to explore the performance of five convolutional network models (resnet, squeezenet, densenet and inceptionv3 and a custom CNN), using Deep Learning Studio: DLS provides GPU training on the cloud; in particular, up to 4 GPUs in its community edition and additional GPUs in its enterprise edition. The DLS models achieved an AUC of 99.77% in detecting cancer cells from the images of cancer cells. Despite the high performances reported in their results, the authors aim to show that the cloud approach can help non-specialists in computer science to exploit melanoma detection issues. In particular, they observed that a common theme in almost all literature contributes is that it is made to appear as a job of specialists in the domain of computers and software engineering.

In [44], Huang et al. exploited the DenseNet CNN model to identify skin cancers and benign skin tumours using KCGMH and HAM10000 datasets. In particular, they aim to build a lightweight skin cancer classification that can be distributed on cloud platforms and mobile devices for remote diagnostic applications. The authors claimed an accuracy reached 89.5% for the binary classifications (benign vs malignant) in the KCGMH dataset; the accuracy was 85.8% in the HAM10000 dataset in the seven-class classification and 72.1% in the KCGMH dataset in the five-class classification.

2.5 evaluation of melanoma classification

In order to evaluate the CNN performance, we used the classical metrics described by the equations below (Equation 2.1- 2.7):

$$\text{Accuracy} = \frac{TP + TN}{TN + FP + FN + TP} \tag{2.1}$$

$$\text{Sensitivity}(TPR) = \frac{TP}{TP + FN} \tag{2.2}$$

$$\text{Specificity(TNR)} = \frac{TN}{TN + FP} \tag{2.3}$$

$$\text{Precision(PPV)} = \frac{TP}{TP + FP} \tag{2.4}$$

$$FDR = \frac{FP}{FP + TP} \tag{2.5}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.6}$$

$$FNR = \frac{FN}{FN + TP} \tag{2.7}$$

TP, TN, and FP are the numbers of correctly predicted true positives and true negatives, whereas FP and FN are the numbers of incorrect predicted false positives and false negatives, respectively. The degree to which the measured value of a quantity corresponds to its true value is known as accuracy (ACC). The sensitivity (SN) of a test refers to its ability to detect true positives. Finally, the ability of a test to detect true negatives is measured by its specificity (SP). It is important to note that in the MIBCP context, we consider the specificity and the FNR, described below, as the primary and most essential metrics due to our goal to identify which technology can minimize the type 2 error.

Precision, also known as Positive Predictive Value (PPV), is a statistical measure that shows the percentages of true positive values in a test. The False Discovery Rate (FDR) measures the frequency of type I errors in null hypothesis testing.

2.5.1 Explainable AI - XIA

Explainable AI (XAI) refers to AI systems that can provide a human-understandable explanation for their predictions, decisions, or actions [35]. The term "explainability" is used to describe the degree to which a person can understand the internal workings of a model or system. The goal of XAI is to create AI systems that are not only effective and efficient but also transparent and understandable to the people who use them.

XAI is becoming increasingly important as AI is being used to make more critical decisions, such as medical diagnoses, financial predictions, and criminal investigations. In these scenarios, it is essential to understand how an AI system arrived at its decision so that the results can be trusted and validated. Furthermore, XAI can help build trust in AI systems, especially among people sceptical of the technology [23]. There are several approaches to achieving explainability in AI, including:

1. Model interpretability: This involves designing AI models that are inherently simple and transparent, such as linear regression models or decision trees. These models are easily understood and easily explained by their structure and weights;

2. Post-hoc explanation: This approach generates explanations for predictions made by complex, black-box AI models, such as deep neural networks. Techniques such as feature importance, saliency maps, and local interpretable model-agnostic explanations (LIME) can be used to generate explanations;

3. Transparency by design: This approach involves incorporating transparency into the design of AI systems from the beginning, such as by using transparent algorithms or creating a system for auditing and explaining AI predictions.

Ultimately, XAI aims to make AI more trustworthy, accountable, and usable. By providing explanations for AI predictions, decisions, and actions, XAI can help build confidence in AI and ensure that it is used responsibly and ethically.

2.5.2 Local Interpretable Model-Agnostic Explanations (LIME)

Local Interpretable Model-Agnostic Explanations (LIME) is a popular open-source software for generating explanations for machine learning models. It is designed to provide human-understandable explanations for the predictions made by complex, black-box AI models, such as deep neural networks [61].

LIME treats a complex AI model as a "black box" and generates an interpretable model local to a specific prediction. This local interpretable model is generated by sampling the input space around the prediction and using the samples to fit a simple, interpretable model, such as a linear regression or a decision tree. The explanations provided by LIME are based on the coefficients of the interpretable model, which can be used to understand how different input features contribute to the prediction.

LIME is "model-agnostic", meaning that it can be used with any machine learning model, regardless of the type of model or the data used to train it: this makes it a flexible and powerful tool for generating explanations for various AI models. LIME can be used in a variety of applications, including:

- debugging and diagnosing machine learning models: By explaining the predictions made by a model, LIME can help identify errors or biases in the model and suggest ways to improve it;
- building trust in AI: By providing explanations for AI predictions, LIME can help build trust in AI systems and ensure that they are used responsibly and ethically;
- improving human-AI interaction: By providing human-understandable explanations for AI predictions, LIME can help to improve communication between humans and AI systems and facilitate more effective collaboration.

Overall, LIME is helpful for anyone interested in developing more transparent and explainable AI systems. Providing local and interpretable explanations for AI predictions can help increase transparency and build trust in AI, making it a valuable tool for practitioners, researchers, and policymakers alike.

3

INCREASING TRUST IN CAD USING FNR MINIMIZATION AND XIA

This chapter compares the main CNN architectures for melanoma detection, highlighting the FNR as the preferred metric and XIA as the way to increase trust in the choices made by CAD. First, the background section discusses the primary issue in early melanoma detection and the advantage of having a low number of false negatives, which is helpful in diagnostic systems (Section 3.1). Then, the used methods are reported (see Section 3.2) with particular attention to dataset preparation and image improvement. Finally, in Section 3.3, we present the results and discuss conclusions 3.4.

3.1 background

Melanoma represents only 1.7% of skin cancers, but it has a high mortality rate due to its capacity to spread fast and metastasize to numerous areas. Therefore, the more effective treatment for prevention is the surgical removal of the primary tumour before tumour cells detach the lymph nodes, allowing the tumour to spread rapidly.

Early detection of melanoma is critical as it considerably reduces mortality in 90% of cases because it will enable therapeutic intervention at a less advanced stage when it is still localized to the site of tumour growth [1]. Furthermore, a study comparing risk-adapted specialized skin surveillance with regular skin screening shows melanomas are more likely to be discovered at an early stage [94].

Unfortunately, populations and screening procedures vary by country, and there are rarely clear criteria. For example, in Germany, regular skin cancer screenings are suggested for people over 35, whereas skin cancer screenings are generally not recommended in the United States. The absence of a standard protocol could lead to a failure in early detection.

Nowadays, the standard way to perform a people check-up comprises a whole-body skin exam, often supported by dermoscopy or other imaging techniques: both exams are performed by an expert, a human. If the expert detects a potential risk naevus, it is mandatory to execute a biopsy to provide a correct melanoma or non-melanoma diagnosis. Unfortunately, this standard way suffers, at least, from the following drawbacks.

The missing of an internationally accepted standard for the screening procedures of melanoma makes it challenging to have standardized data sets that can help with statistical and exploratory analysis. Also, a manually performed body scan can be slow because the dermatologist often needs to use the dermatoscopic. Interestingly, even though clinical images are easy to capture and could provide similar performance as dermoscopic images [13], most of the works in the literature use dermoscopic images. In the Internet of Things (IoT) and Internet of Medical Devices (IoMD) era, CAD system services should be provided to patients without needing to visit the clinic physically or have a dermoscopic at home: We are moving toward digital health. Therefore, we expect that patients should be able to take a photo, send it to web services, and receive preliminary answers regarding whether they should be seen or not.

Another issue is related to human skills: the screening is done by one or more human experts who rely on their skills and knowledge: If something goes wrong with these skills, a biopsy may be requested for a simple naevus, leading to an invasive operation, the biopsy, for the patient. On the contrary, a biopsy may not be requested for melanoma. In that case, we could have type 1 (false positive) and type 2 (false negative) classification errors. A type I error occurs during verifying a statistical hypothesis when the true null hypothesis is incorrectly rejected. A type II error is the failure to reject an incorrect null hypothesis. Following that, the False Positive Rate (FPR) and False Negative Rate (FNR) can be defined as the proportion of all negative results that lead to positive test outcomes and the proportion of positives that lead to adverse test outcomes, respectively.

Fortunately, AI has shown the potential to outperform dermatologists in dermoscopic melanoma diagnosis [72]. In addition, the CNN has been shown to provide the most accurate and precise results for constructing skin lesion classifiers [4]-[26]: the significant improvement made by these results is that unnecessary biopsies are frequently avoided while needed biopsies are missed only a few times; this significantly reduces FNR and FPR.

Then, the most used CNNs architecture performance was studied to identify the best network, in terms of FNR, that could be eligible for a CAD in the melanoma detection field using clinical images. Furthermore, we have chosen to consider FNR minimization because missing a needed biopsy on the skin is more dangerous (life-threatening) than making a biopsy without melanoma.

The neural networks studied, updated and trained are AlexNet [54], DenseNet [43], Google Inception V3 [87], GoogleNet [88], MobileNet [41], ShuffleNet [102], SqueezeNet [45] and VGG [86].

3.2 methods

3.2.1 Dataset preparation

In this work, the dataset presented in developing the MED-NODE computer-assisted melanoma diagnosis system, called in this document MED-NODE [33], was used as the primary image source. MED-NODE is composed of 70 images of melanoma and 100 images of naevi. All the images were clinical. Then it was taken without the utilization of a dermoscopic.

Due to the small dataset size, multiple combinations of image operators were applied to the original dataset: data augmentation (DA) and image optimization to extend the training dataset size. DA can aid in the extension of small datasets and the improvement of prediction performance.

In particular, three new training sets were built from the original MED-NODE dataset by applying different image operations and DA operators to the same images differently.

The operators used to perform data augmentation to build the new dataset were: random rotation, random scaling, and random translation on X and Y. With these operators, we built a new dataset named “NSA" containing the MED-NODE original images and the new images generated by the DA operations applied to the MED-NODE original images.

The DA performed can be reproduced by using the following code (with theimageDataAugmenter object provided by MATLAB):

```
imageAugmenter = imageDataAugmenter( ...
                 'RandRotation',[-180,180], ...
                 'RandScale',[1,100],...
                 'RandXTranslation',[-180 180], ...
                 'RandYTranslation',[-180 180])
```

Using NSA, we could compare the Neural NetworK (NN) performance to understand how data augmentation impacts NN prediction performances in this specific case of MCIBCP. The results of the comparison are available in the following sections. We were also interested in evaluating the impact of the image quality improvement techniques on NN classification performances; in particular, we used the pre-processing quality step (IIQ) and a simple segmentation process (OTSU). More details regarding these two techniques are available in the next subsection.

From the original MED-NODE dataset, we built the following new datasets:

- “INA",which contains MED-NODE original images improved by combining IIQ and the OTSU method (IIQpOTSU);
- “IA", which contains NSA images improved by combining IIQ and the Otsu method (IIQpOTSU).

For clarity, the acronyms used to identify each dataset can be interpreted as:

- “INA", containing MED-NODE original images by using quality improved and data augmentation techniques;
- “NIA", containing MED-NODE original images not quality improved but using data augmentation techniques;
- “IA", containing the NSA images by using quality improved and data augmentation techniques;

For coherence, the original MED-NODE dataset was renamed into NINA (Not improved, Not Data Augmented) in the following sections.

3.2.2 Image improvement method

Clinical images often suffer from poor contrast: to improve the quality of the MED-NODE, a MATLAB routine was used: this routine executes histogram optimization and enhances the contrast of coloured images. Image enhancement means improving an image's perceptibility so that the final product is superior to the original: Image contrast enhancement before further preprocessing can improve analysis results [71]. The function used for implementing the IIQ operation is listed below:

```
data = function IIQ(data)
s\_l = rgb2lab(data);
max\_l = 100;
L = s\_l (:,:,1)/ max\_l;
sh\_a = s\_l;
sh\_a (:,:,1) = adapthisteq(L)* max\_l;
sh\_a =  lab2rgb(sh\_a);
data= sh\_a;
```

In Figure 3.1 and Figure 3.2, a sample image before (a) and after the IIQ application (b) is shown for naevi and melanoma images, respectively.

3.2.3 Image segmentation method

In order to investigate how segmentation might impact training performance, the OTSU method for the segmentation process was used. In particular, OTSU was used to make two of the three datasets described in the next section. OTSU performs automatic image thresholding, separating the pixels into background and foreground [66]. OTSU segmentation was performed using the following code:

```
[input_image,map] = imread(F);
bw_input = rgb2gray(input_image);
[T, EM] = graythresh(bw_input);
BW = imbinarize(bw_input, T);
mask_otsu = BW;
mask_otsu= ~mask_otsu;
new_image = input_image * mask_otsu;
```

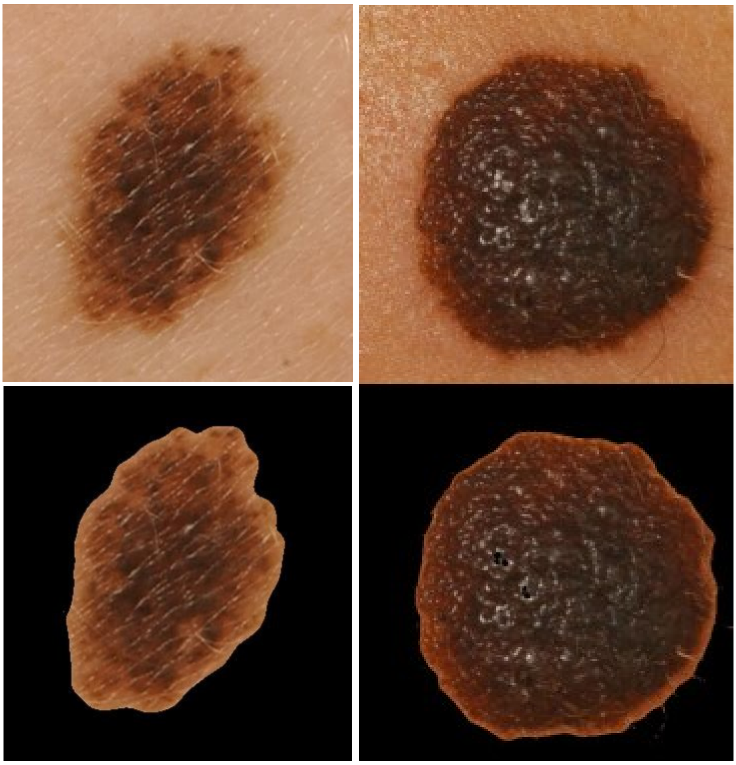


Figure 3.1: Images of Naevi before and after application are shown from top to bottom. Specifically in the upper part the images before the application where you can also see the skin while in the lower part the images after the segmentation.

3.2.4 CNN refactoring and evaluation

In order to identify the CNN that ensures the best FNR regarding the MCIBCP, the following networks were refactored: Alexnet, DenseNet, GoogleNet Inception V3, GoogleNet, MobileNet, ShuffleNet, SqueezeNet and VGG16. The performances of these network in terms of FNR was analyzed.

The original version of CNN comes with pre-trained weights to solve a multi-class classification problem. In particular, these networks were trained on ImageNet [55] and can discriminate between 1000 classes of objects. All the pre-trained weights from these networks were discarded as a first step. These preliminary steps removed all possibilities of transfer learning from the Ima-

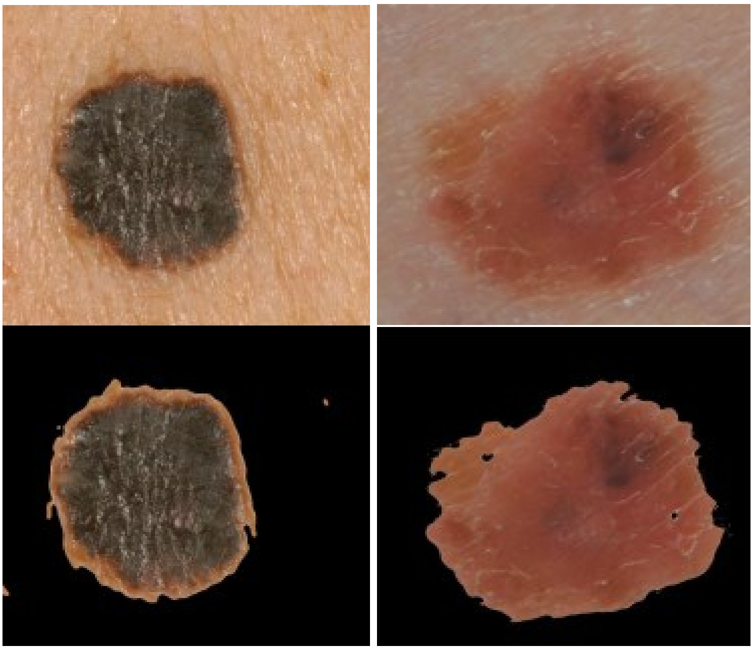


Figure 3.2: Images of Melanoma before and after application are shown from top to bottom. Specifically in the upper part the images before the application where you can also see the skin while in the lower part the images after the segmentation.

geNet upon which all these networks were pre-trained. Also, all the final layers (softmax, Fully connected) of all the CNNs were changed to allow these networks to discriminate between two classes instead of one thousand classes.

In [21], the results reported strongly suggested that the training and validation steps could suffer from intra-class dissimilarities and extra-class similarities. In particular, we rely on the hypothesis that the CNN performances can vary, even if the training, validation, and test sets vary minimally. This fact can be observed in [34] when the ISIC 2018 winning algorithms performances dropped to a coin flip performance by only adding a new object class.

In order to avoid biased results, a similar training protocol used in [21] was followed to consider the mean performance instead of absolute performance to make our analysis more robust. The experimental environment used was MatLab 2021b[1].

One hundred training steps were performed for each network and dataset: 3,200 (training, validation, and test) steps were performed to collect the experimental data. Each training step was performed by using `MaxEpochs=30`, `MiniBatchSize=32` and `InitialLearnRate=1e` ^{4}. For each training step, the training, validation, and test sets were allowed to change slightly while the previous network weights were discarded. No transfer learning was allowed during the training steps. The dataset was divided using the following ratios for each iteration: 0.5 for the training set (85), 0.3 (51) for the validation set, and 0.2 (34) for the test set. The randomized option of `splitEachLabel` method was enabled. The training set was split equally between melanoma and non-melanoma photos, chosen randomly from the starting image collection for each cycle. In addition, each image was resized to fit the network's input constraints. For example, for Google Inception V3, the input images were resized to 299x299, while for AlexNet, the images were resized to 227x227 pixels. The training and validation sessions were executed using the `trainNetwork` function, while the `classify` function executed test sessions.

3.2.5 XIA analysis

In order to allow physicians to evaluate network results, LIME was used. In particular, for each classification result, an image showing the naevus section that allowed the NN to determine the choice between benign or malign classes is generated. Figure 3.3 reports an example of the LIME output on the results obtained using InceptionV3 NN. The classificator and the LIME output were analysed in collaboration with the Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana" (DIPMED).

Due to hardware limitations, LIME was executed only on Clinical Images in this work. Also, preliminary tests were executed using Colab. The preliminary results suggested that the NN trained also uses the background to make the prediction. Conse-

[1] https://www.mathworks.com/products/matlab.html

quently, the segmentation pre-processing step must be improved to avoid noise in the background.

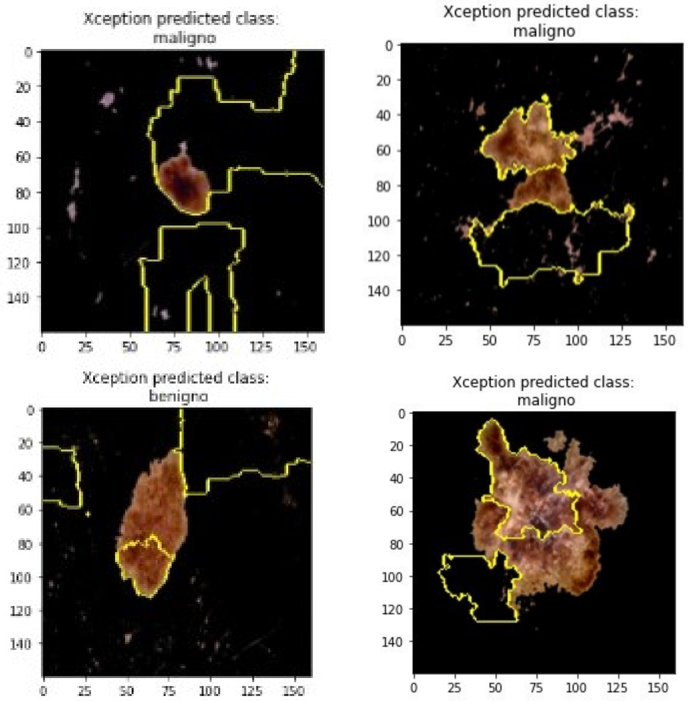

Xception predicted class:
maligno
Xception predicted class:
maligno
Xception predicted class:
benigno
Xception predicted class:
maligno
0
20
40
60
80
100
120
140
0
25
50
75
100
125
150

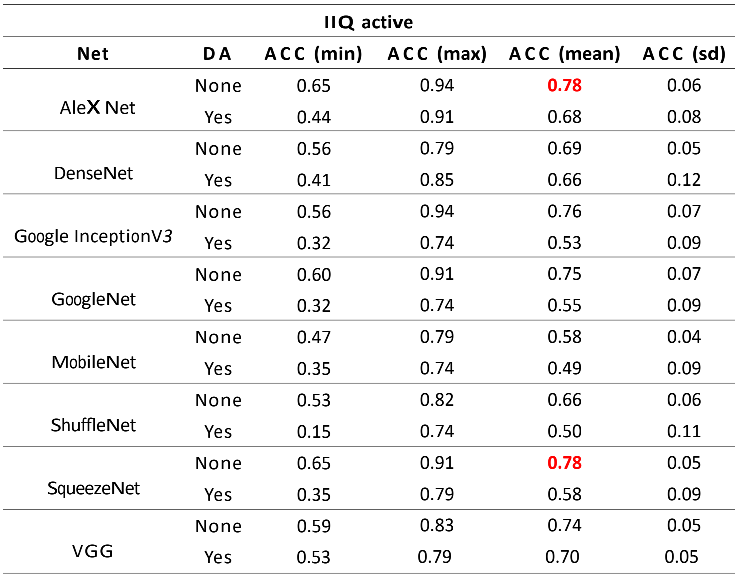
Figure 3.3: The InceptionV3 prediction explained by LIME.

3.3 results and discussion

In this section, the results collected in each experiment are presented in terms of average, maximum, minimum, and standard deviation values for the ACC to identify which CNN performs better globally. In addition, the results are presented in a form emphasising the importance of having the lowest FNR possible in early melanoma detection.

Table 3.1 reports the accuracies of all the CNNs using the “IA" and “INA" datasets. For each CNN, two rows are used to show accuracies when data augmentation is used or not, while imaging optimization techniques are always used. The best mean accuracy results for the AlexNet and SqueezeNet networks are highlighted in red, settling at 78%. These findings highlight how these two CNNs might resist inter-class/extra-class issues most. Interestingly, Google InceptionV3, GoogleNet, and VGG reach an average accuracy greater than 70% when the data augmentation is not used on the INA dataset. Overall, all tested CNNs perform better on the INA dataset, suggesting that data augmentation using scaling, rotation, and translation may reduce classification performance. Again, AlexNet obtained the best global performance.

Table 3.1: The global performances of the CNNs on the INA and IA datasets are reported. In addition, image improvement techniques are active.



IIQ active					
Net	DA	ACC (min)	ACC (max)	ACC (mean)	ACC (sd)
AleX Net	None	0.65	0.94	**0.78**	0.06
	Yes	0.44	0.91	0.68	0.08
DenseNet	None	0.56	0.79	0.69	0.05
	Yes	0.41	0.85	0.66	0.12
Google InceptionV3	None	0.56	0.94	0.76	0.07
	Yes	0.32	0.74	0.53	0.09
GoogleNet	None	0.60	0.91	0.75	0.07
	Yes	0.32	0.74	0.55	0.09
MobileNet	None	0.47	0.79	0.58	0.04
	Yes	0.35	0.74	0.49	0.09
ShuffleNet	None	0.53	0.82	0.66	0.06
	Yes	0.15	0.74	0.50	0.11
SqueezeNet	None	0.65	0.91	**0.78**	0.05
	Yes	0.35	0.79	0.58	0.09
VGG	None	0.59	0.83	0.74	0.05
	Yes	0.53	0.79	0.70	0.05

Table 3.2 reports the accuracies of all the CNNs using the “NINA" and “NIA" datasets. Like Table 3.1, two rows are used for each CNN to show accuracy when data augmentation is used. In this case, no IIQ techniques are active.

AlexNet reaches a mean accuracy of 89% when no data augmentation is used. GoogleNet settled on 80%. Again, the results suggest that the best outcomes for all networks can be obtained without data augmentation techniques.

In Figure 3.4, a summarization of the results regarding global performance is reported. The red box highlights that AlexNet performs best in the four conditions (IA, INA, NIA, NINA).

Table 3.2: The global performances of the CNNs on the NINA and NIA datasets are reported. In addition, image improvement techniques are not active.

		IIQ not active			
Net	**DA**	**ACC (min)**	**ACC (max)**	**ACC (mean)**	**ACC (sd)**
AleX Net	None	0.68	**1**	**0.89**	0.05
	Yes	**0.76**	0.97	0.87	0.05
DenseNet	None	0.62	0.79	0.74	0.04
	Yes	0.41	0.88	0.73	0.08
Google InceptionV3	None	0.56	0.94	0.74	0.07
	Yes	0.32	0.71	0.55	0.07
GoogleNet	None	0.65	0.94	0.80	0.06
	Yes	0.30	0.76	0.55	0.09
MobileNet	None	0.50	0.91	0.75	0.09
	Yes	0.35	0.76	0.56	0.08
ShuffleNet	None	0.44	0.88	0.69	0.08
	Yes	0.26	0.74	0.52	0.10
SqueezeNet	None	0.38	1	0.55	0.11
	Yes	0.15	0.79	0.58	0.10
VGG	None	0.59	0.82	0.75	0.04
	Yes	0.59	0.82	0.73	0.05

Table 3.3 and Table 3.4 report the standard metrics for evaluating the tested networks: The first one reports the results for the “IA" and “INA" datasets; the second one reports the results for the “NINA" and “NIA" datasets. The table structure is the same as the previous two, with two rows for each network: Figure 3.5 and Figure 3.6 report the table data graphically; each column’s colours are identified with an acronym with the form metric_[DA]; as an example, SN_DA means sensitivity with data augmentation, if _DA is omitted, the metric value is considered in the case without data augmentation. As expected, the best results for SP and SN are obtained without data augmentation due to the previous accuracy results presented: we can see that experiments without data augmentation outperform all methods except SqueezeNet, which is the only exception; in SqueezeNet, SP and SN values increase with data augmentation.

However, in the context of MCIBCP, the FNR takes on more weight because it is directly related to the type 2 error. Therefore, the results reported in Table 3.3 and Table 3.4 suggest:

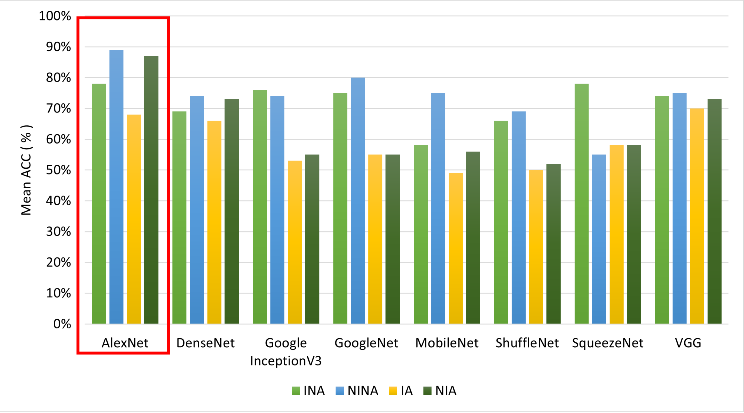


Figure 3.4: The global performances of the CNN on the four datasets.

- SqueezeNet ensures the lowest FNR (0.13) on the “INA" dataset;
- AlexNet ensures the lowest FNR when used on the “NINA" dataset (0.13);
- DenseNet ensures the lowest FNR on the “IA" dataset (0.27);
- VGG ensures the lowest FNR on the “NIA" dataset (0.07).

Interestingly, even though SqueezeNet is confirmed as the worst network in global terms, it ensured the lowest FNR in at least one case. Therefore, SqueezeNet in the INA situation can be chosen to minimize type 2 errors.

3.4 conclusion

Melanoma is a severe type of skin cancer responsible for about 99,780 new malignant diagnoses[2]. However, melanoma can be cured in most cases with an early diagnosis. Therefore, early diagnosis is critical in this context. Melanomas exist in many different shapes, sizes, and colours and affect people with all skin types. Dermatologists use these characteristics to apply the ABCDE rules that can be used to estimate the degree of threat

[2] https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

Table 3.3: The FNR and the other metrics of the CNNs on the IA and INA datasets are reported. The image improvement techniques are active.

IIQ active							
Net	**DA**	**SN**	**SP**	**PPV**	**FDR**	**FNR**	**FPR**
AlexNet	None	0.75	0.82	0.73	0.27	0.25	0.18
	Yes	0.63	0.76	0.65	0.35	0.37	0.24
DenseNet	None	0.51	0.81	0.78	0.22	0.34	0.19
	Yes	0.55	0.69	0.62	0.38	0.27	0.24
Google InceptionV3	None	0.74	0.79	0.68	0.32	0.26	0.21
	Yes	0.38	0.57	0.38	0.62	0.47	0.41
GoogleNet	None	0.72	0.78	0.67	0.33	0.28	0.22
	Yes	0.44	0.62	0.40	0.60	0.51	0.37
MobileNet	None	0.37	0.59	0.09	0.91	0.48	0.41
	Yes	0.38	0.51	0.59	0.41	0.53	0.32
ShuffleNet	None	0.55	0.75	0.67	0.33	0.41	0.25
	Yes	0.39	0.57	0.51	0.49	0.55	0.42
SqueezeNet	None	0.32	0.59	0.79	0.21	0.13	0.11
	Yes	0.39	0.63	0.42	0.58	0.42	0.36
VGG	None	0.59	0.80	0.71	0.30	0.26	0.20
	Yes	0.61	0.65	0.64	0.36	0.30	0.18

regarding a naevus. Unfortunately, nowadays, the last word regarding the malignancy of a lesion is delegated to the biopsy, which performs the histological analysis of the suspected lesion. Unfortunately, this state-of-the-art protocol can lead to delays in diagnosis and unnecessary invasive surgery in the case of FP. In the case of non-detection of melanoma, this FN outcome could result in potentially fatal circumstances. In recent years, multiple computer-aided diagnoses (CAD) systems working on melanoma images have been proposed to speed up diagnosis. In addition, some results in the literature suggest that artificial intelligence techniques can outperform dermatologists in melanoma diagnosis, particularly CNN. These networks have been proven to give the most accurate and exact results for choosing between benign and malignant outcomes. If the accuracy of these CNN continues to grow in the future, unnecessary biopsies (type 1 error – FP)

Table 3.4: The FNR and the other metrics of the CNNs on the NINA and NIA datasets are reported. The image improvement techniques are active.

IIQ not active							
Net	**DA**	**SN**	**SP**	**PPV**	**FDR**	**FNR**	**FPR**
AlexNet	None	0.87	0.90	0.86	0.15	0.13	0.10
	Yes	0.84	0.91	0.87	0.14	0.16	0.09
DenseNet	None	0.56	0.82	0.77	0.23	0.29	0.18
	Yes	0.64	0.74	0.56	0.44	0.19	0.26
Google InceptionV3	None	0.73	0.76	0.62	0.38	0.27	0.24
	Yes	0.39	0.60	0.29	0.71	0.57	0.40
GoogleNet	None	0.79	0.82	0.72	0.28	0.21	0.18
	Yes	0.45	0.63	0.48	0.52	0.54	0.37
MobileNet	None	0.81	0.72	0.45	0.55	0.14	0.28
	Yes	0.32	0.61	0.29	0.71	0.37	0.38
ShuffleNet	None	0.61	0.74	0.60	0.40	0.35	0.26
	Yes	0.36	0.60	0.45	0.55	0.52	0.39
SqueezeNet	None	0.23	0.43	0.43	0.57	0.22	0.27
	Yes	0.43	0.62	0.41	0.59	0.41	0.37
VGG	None	0.58	0.83	0.76	0.24	0.27	0.17
	Yes	0.82	0.59	0.40	0.60	0.07	0.24

will be avoided more and more, while needed biopsies (type 2 error – FN) will be missed only a few times.

In this complex context, where early melanoma treatment, minimizing FNR and providing easy-to-use tools to physicians is critical, our work aims to investigate the current CNN architectures available. In particular, we aimed to identify the CNN network structure that ensures the lowest FNR when used with Clinical Melanoma Images: nine CNNs, including Alexnet, DenseNet, GoogleNet Inception V3, GoogleNet, MobileNet, ShuffleNet, SqueezeNet, and VGG16 were evaluated. We started from the MED-NODE dataset, which includes 170 clinical photos (70 images of melanoma and 100 images of naevi) extracted from the digital image archive of the Department of Dermatology of the University Medical Center of Groningen (UMCG). Due to the small size of the dataset, we used image improvement and data augmentation techniques; four datasets (NINA, NIA, INA,

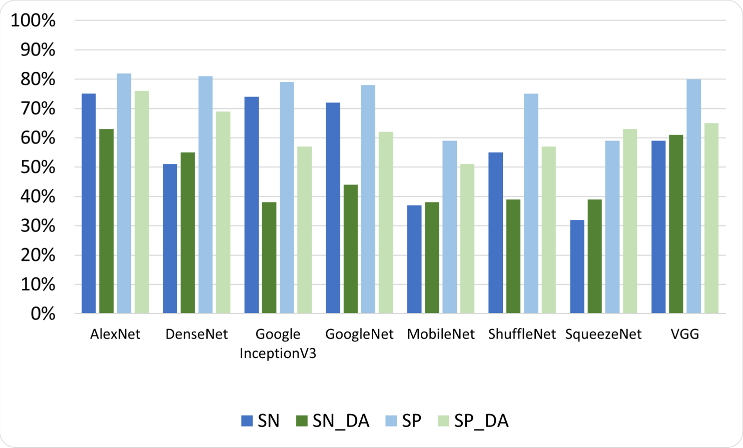


Figure 3.5: Comparison of SN and SP for the INA e IA datasets.

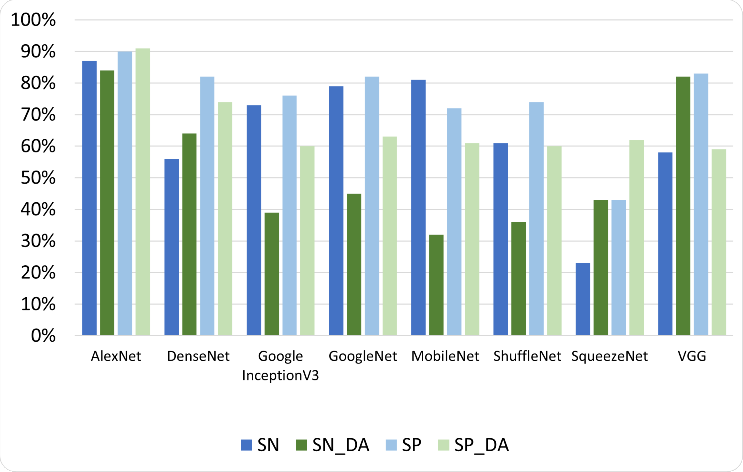


Figure 3.6: Comparison of SN and SP for the NINA e NIA datasets.

IA) were generated to investigate the impact of DA and image preprocessing on the final classification performance. The training, validation, and test sessions were executed on each dataset. Overall, all tested neural networks, with one exception, perform better without data augmentation, with a maximum accuracy of 0.78% achieved by AlexNet and SqueezeNet. In the absence of preprocessing and data augmentation, AlexNet performed best with 0.89%, 0.75% and 0.82% of accuracy, sensitivity, and specificity, respectively. In the context of MCIBCP, however, the

FNR is more important than global accuracy because it is directly related to type 2 errors, which can result in life-threatening situations. The results suggest that the VGG CNN can ensure the lowest FNR at the expense of global accuracy, while AlexNet can ensure comparable FNR like VGG but with the highest global accuracy. Therefore, VGG and AlexNet were the CNNs that might be used to build a CAD system, FNR-driven and easy to use due to the capability to use clinical images instead of dermoscopic images. The remarkable results obtained with clinical images alone, whose quality is unquestionably lower than that of dermoscopic images, enable help in prevention in a situation where it is crucial.

Finally, the results support what has already been discovered: that networks perform better when using the original images without any preprocessing [19, 21]. Additional research on this aspect might aid in understanding the motivation behind this behaviour. Furthermore, future research could investigate local and global features relevant to melanoma, other neural networks, and different image preprocessing techniques to minimize the FNR while maximizing global accuracy and other metrics.

4 GENETIC ALGORITHMS FOR MELANOMA CLASSIFICATION

This Chapter proposes and describes the contribution to the scientific literature for diagnosing melanoma using genetic algorithms, which represents the main idea of our work (see Section 4.1) . The Chapter will concentrate on two key topics: the use of ge-netic algorithms in clinical images and the application of genetic algorithms to clinical and dermoscopic images.

The first work, detailed in Section 4.2, has a vital position in Smart healthcare, contributing to the creation and growth of the Internet of Medical Things (IoMT). For example, in future, thanks to systems capable of recognizing melanoma from clinical images, patients might be able to conduct an increasingly accurate monitoring process using the devices at their disposal.

Prevention in the field of melanoma identification is essential for lowering the mortality rate and enhancing patient lives. Therefore, we intend to present a preliminary approach to the detection of melanoma through clinical imaging in our initial work. In the second study, presented in Section 4.3, we aimed to expand and enhance the prior contribution by focusing on dermoscopic images to compare results with clinical images to contribute to and facilitate the dermatologist's work.

The contributions presented below add significant novelty to the literature through the innovative use of genetic algorithms. Evolutionary algorithms, also known as genetic algorithms, optimize deep learning models' parameters (tuning) using the biological concept of evolution and its operations [98]. However, the work presented uses the GAs for network structure optimization rather than parameter optimization to build new and efficient networks using the evolved processes.

4.1 idea

A GA is an optimization method influenced by the biological principles of natural selection and evolutionary theory [53]. It is used to discover the optimal solution to an optimization problem by simulating natural Selection. John Holland introduced genetic algorithms in 1975, based on Darwin's evolutionary theories outlined in his 1859 book, "On the Origin of Species by Means of Natural Selection and the Preservation of Favoured Races in the Struggle for Life".

These algorithms imitate the processes of evolution by conforming to the Darwinian theory that the elements of the environment with the most significant potential for adaptation have a greater chance of surviving and passing on their traits to subsequent generations. Therefore, there is a population of individuals, each of which has n chromosomes and was created at random. These individuals continue to change from generation to generation by mechanisms analogous to the natural process of evolution. Chromosomes almost always take the form of binary strings when they are stored. Each locus (a particular location on a chro-mosome) is composed of two alleles (various versions of genes) represented by the numbers 0 and 1 [56].

More in general, the goal of a GA is to build a population of candidate solutions to a problem, evaluate each candidate solution using a fitness function (scoring function), and then select the best solutions to generate a new population. This method is carried out iteratively until either a solution considered adequate is discovered or a particular stopping requirement is satisfied. Next, the Selection of solutions for raising the future generation is based on their fitness, with the most likely candidates being the most effective. Finally, the new population is produced using a process known as crossing and mutation. Then, two-parent solutions are united to produce a child solution, and random alterations are introduced to expand the population's genetic diversity. This process continues in an iterative way, which ultimately results in the enhancement of solutions through time. To be more explicit, the following is a list of the basic operations that are involved in a genetic algorithm:

1. **Initialization**: The first step is randomly populating the population of potential solutions.

2. **Selection**: In this phase, the GA selects the most suitable parents from the present population. There are numerous selection methods, including Selection by the roulette wheel, Selection by tournament, and Selection by rank.

3. **Crossover**: This is the process of recombining selected parents' genetic material to produce offspring solutions. The progeny solutions will inherit qualities from both parents.

4. **Mutation**: Mutation is the process of introducing random changes to the genetic material of progeny, helping prevent the GA from becoming locked in a local optimum by introducing new genetic variations into the population.

5. **Evaluation**: This step evaluates the fitness of each solution in the population. The fitness function assigns each solution a score based on how well it solves the current problem.

6. **Replacement**: The GA is responsible for replacing the current population with a new generation of solutions. This process is repeated until a satisfactory solution is found or for several generations.

In each generation, these operations are repeated until the termination criteria are met. The criteria could be based on the number of generations, the best solution's fitness, or other factors [57].

GAs have numerous applications in different fields, including optimization problems in engineering, finance, and machine learning. They are especially suited to problems that are difficult to solve using traditional optimization techniques, such as problems with an ample search space, nonlinear constraints, and multiple local optima [49]. Although genetic algorithms have been utilized successfully in a variety of fields, they can be computationally costly due to the training and evaluation of a large number of candidate solutions over successive generations [101]. In addition, the optimization quality depends on the fitness function and algorithm parameter values, which can be challenging to alter. However, genetic algorithms can be a helpful tool for exploring many solutions to a given optimization problem and supplement more conventional optimization techniques like gradient-based optimization [24].

GA and DL are two independent disciplines of research. However, they can be merged for specific purposes. It helps optimize various components of deep learning models, including optimizer settings, model hyperparameters, and even the architecture of networks [12].

Inside deep learning, a genetic algorithm can tune deep learning model hyperparameters such as the number of hidden layers, neurons per layer, learning rate, and smoothing coefficient. In this case, the deep learning model can be used as a fitness function to assess the performance of the genetic algorithm's candidate solutions. The genetic algorithm can then tune the hyperparameters of the deep learning model to achieve the best performance on a given task, such as image classification or speech recognition [98].

Combining genetic algorithms and deep learning has additional applications in the evolution of neural network architectures [84]. In this case, the genetic algorithm can be employed to develop new neural network structures and evaluate their performance on a particular task, with the best-performing architectures providing the foundation for the subsequent generation. Furthermore, this may enable the discovery of creative neural network topologies outperforming conventional hand-designed models. The latter is the case study developed in the two works that will be discussed in the following subsections.

4.2 genetic algorithms on clinical image

This contribution discusses the initial findings we achieved by merging the main characteristics of genetic algorithms (GA) with the convolutional neural network (CNN) in order to address the melanoma detection challenge (GACNN).

The MED-NODE clinical images dataset was used as the data source. In addition, the capabilities of GACNN and AlexNet, both with and without Otsu segmentation, were investigated. In addition, the textitaccuracy was used as the scoring function for the GA evolution process. This work suggested that the proposed method could improve melanoma categorization by enabling the network design to develop independently of patient involvement.

The following sections report an overview of the dataset utilized in subsection 4.2.1. Also, the description of the methodology is reported in subsection 4.2.2. Finally, the results are presented by discussing the relevant findings in subsection 4.2.3. In the subsection 4.2.4, the implications of these findings on the direction of upcoming developments are discussed.

4.2.1 Dataset

The MED-NODE melanoma images dataset (MED-NODE) is a specialized subset of the MED-NODE medical knowledge graph (MNMKG) that primarily focuses on melanoma-related information. The MNMKG contains images and melanoma information, including the disease's origins, symptoms, diagnostic procedures, and potential treatments. Also, the MNMKG graph represents entities such as diseases, medications, and treatments.

In particular, the edges are used to describe the interactions between the entities. For instance, the dataset may contain information regarding the relationship between a specific type of melanoma and the risk factors that are associated with it, or the information may concern the relationship between a treatment and the adverse effects that are associated with it. This dataset can assist researchers and medical practitioners in improving patient outcomes by assisting them in making more informed decisions and giving a comprehensive and structured representation of the knowledge regarding melanoma. For this study, only the images contained in MNMKG were used.

Therefore, the following section identifies this particular subset of MNMKG with the acronym MED-NODEs: the used dataset consists of 170 clinical images and includes, among other things, 70 examples of melanoma and 100 examples of benign nevi [33].

4.2.2 Materials and Method

In the environment where we conducted our experiments (`Matlab 2021`), we defined our working objects using the nomenclature used in GA. If this notation $F(t)$ were utilized, it would denote the composition of an object at the moment t. To be more specific, we define an entity E_i as a vector E_i = {F_1, . . . , F_m}, consisting of m features. We referred to each characteristic or feature of a generic entity as a gene of that entity. We called each feature F_j of a generic entity E_i a gene of E_i. The entire set of genes is called the Genome of E_i. Within the context of our simulation, a gene may stand in for a Matlab CNN core object (network layer) or a pre-processing routine, such as Otsu [75]. Each characteristic Fj may or may not be expressed by E_i: this means that a new entity E_k could inherit a gene F_e from an existing entity E_i that has begun expressing it; consequently, a new entity Ek could inherit a gene Fe from an existing entity Ei that has begun expressing it. The set $P(t) = \{E_1, \ldots, E_n\}$ is called Population at time t. The population size n(t) may change following the time evolution identified by t.

The following constraints were defined:

1. The initial gene of each entity must be an image input layer and
   or one of the previously defined pre-processing routines.

2. If the gene q is a pre-processing routine, then the gene g + 1 must be an image input or another pre-processing layer.

3. The final gene of an entity must be a classification layer.

In the context of this experiment, the population refers to the collection of all living entities. For the experiment, we restricted the gene types that an entity was allowed to utilize to the following: "Convolution," "ReLu," "Cross Channel Normalization," "Max Pooling," "Grouped Convolution," "Fully Connected Layer," "Dropout," and "Softmax.".

At each stage of evolution, all entities whose genes have been expressed but are incompatible with the environment are quickly destroyed. The entity will perish at the first step if it exposes a gene pipeline that Matlab's focus training function does not permit; this contradicts the terms of the function. For example, if the first gene of the entity E_i is a II layer with input dimension D = Width ⇥ Height ⇥ Depth and the second gene is a C layer, it must use the same D input size. If this does not occur, the training function will fail, and we will assume that the gene as it is expressed is incompatible with the environment that is being simulated (melanoma classification).

The following setup was utilized during the training of each compatible entity:

```
('sgdm', ...
'MaxEpochs', 16, ...
'MiniBatchSize', 12, ...
'Shuffle', 'every-epoch', ...
'InitialLearnRate', 0.0001, ...
'ExecutionEnvironment', 'auto')
```

The function that drove population evolution was the maximization of global population accuracy. For each evolutionary step, the maximum accuracy of each survivor was calculated. Therefore, all entities that reveal an accuracy at time t equal to or greater than the highest accuracy achieved by the generation that came before it, t 1, will survive to the next generation.

In addition, 10% of entities randomly picked survives in each evolution step regardless of the accuracy exposed at time t. The GA was terminated if there was no apparent gain in accuracy after ten iterations of the evolutionary process. Due to the limitation of the cloud platform used, the amount of feasible crossover and mutation was limited.

Consequently, each surviving organism was restricted to ten mutations and one hundred crossovers. We attempted to work around these limitations by using a randomized population of 10,000 entities as our starting point. We attempted to alleviate these limitations by generating a random population of 10,000 entities.

4.2.3 Results and Discussion

After running the AlexNet network both with and without the Otsu segmentation applied to MED-NODE, we carried out the training procedure one hundred times. After that, we ran the GACNN, enabling the system to evolve for one hundred iterations.

Accuracy was the criterion employed for our reference focus (ACC): the average ACC (which is denoted by the "mean ACC"), maximum ACC (which is denoted by the "max ACC"), minimum

ACC (which is denoted by the "min ACC"), and Standard Deviation (which is denoted by the "SD") for the standard AlexNet execution were computed, as shown in Table 4.1. The average AlexNet performance was 0.81%, while the best AlexNet performance was 0.97%. Before completing the one-hundredth iteration, the maximum ACC achieved with GACNN is 0.97.

MED-NODE

Net	Segmentation	min ACC	max ACC	mean ACC	SD
AlexNet	-	0.68	0.97	0.81	0.06
	Otsu	0.50	0.91	0.72	0.07
GACNN	-	0.68	0.97	-	-

Table 4.1: Performance of AlexNet on the MED-NODE dataset.

The preliminary findings of this work indicate that GAs may be able to direct the construction of a neural network structure with performance comparable to that of traditional neural network training methodologies. In addition, the preliminary findings of our research indicate that the NN design discovered by GACNN is likely to be more stable than the conventional CNN, which in this instance, is AlexNet. We found that GACNN had a higher average accuracy than AlexNet (calculated over 100 runs), which decreased the average performance hit that was brought on by changes to the data set that were relatively insignificant. GACNN has a significantly better average accuracy than AlexNet's mean ACC, which is significant.

Figure 4.1 also depicts a collection of plateaus demonstrating a diminishing tendency over an average of nine iterations. These results suggest that the population is getting closer to finding the best option. On the other hand, we discovered a significant "birth-death ratio" (up to 95% at each stage of evolution). This observation may suggest that the initial population or recombination stages require more explicit definitions in order to be adequately described.

Finally, the times needed for the executions are rather lengthy. The current stage of the experiments clearly shows that execution times increase proportionally with the complexity of the network,

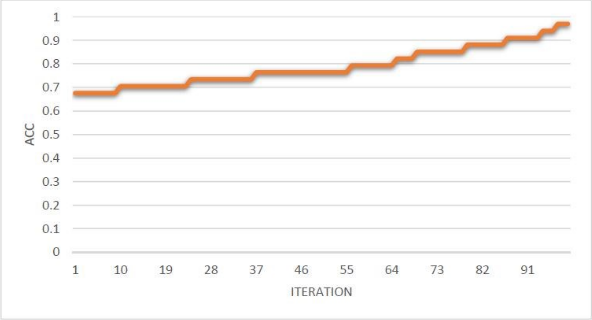


Figure 4.1: GACNN performance over 100 iterations.

related to the chromosome length. The training of each network typically takes about eight minutes. As a result, the entire project was distributed on the cloud.

4.2.4 Conclusion

According to the preliminary findings, using GA to define NN structure design might enable performance levels comparable to those achieved by traditional NN training methods. Specifically, GACNN outperforms AlexNet regarding mean ACC over a hundred executions. A plateau set is also depicted in Figure 4.1, demonstrating a decreasing trend in mean value after every nine rounds. The conclusion that can be drawn from these results is that the population is moving progressively closer to the best solution. On the other hand, we found a significant death ratio (up to 95 per cent for each step of the evolution process). This observation may imply a need for a more precise definition of the processes involving the original population or recombination.

A more in-depth study is required to properly investigate the dynamics of population change and behaviour, particularly the birth-to-death ratio. In addition, the proposed approach can, in the future, be used for additional melanoma datasets (for example, clinical, dermoscopic, or histological) with the help of other cutting-edge evolutionary optimization algorithms.

4.3 genetic algorithms on dermoscopic images

In this section, another convolutional neural network architecture that employs evolutionary algorithms in its design is presented. Like the first approach, the goal is to identify the optimal neural network structure for improved melanoma classification, in that case using dermoscopic images instead of clinical images.

A refined subset of pictures from ISIC, one of the most used datasets for melanoma classification, was employed in the experimental study. The genetic algorithm for building the convolutional neural network enables the population to achieve optimal fitness across successive generations. Preliminary results indicate that the suggested strategy could improve the categorization of melanomas by reducing the necessity for user input and avoiding a priori network design selection.

Subsection 4.3.1 discusses the dataset used, whereas subsections 4.3.2 and 4.3.3 detail its preprocessing and image modelling for training the networks. Next, in subsections 4.3.4 and 4.3.5, the definition, parameters, heuristics, and fitness function utilised for GAs are presented in depth. Then, subsection 4.3.7 will offer the results with a discussion of their significance and a comparison to the relevant literature. Last but not least, subsection 4.3.9 finishes our contribution by discussing the obtained results and future potential.

4.3.1 Dataset

The International Skin Imaging Collaboration (ISIC) is an academic-industry collaboration that aims to make it easier to use digital skin imaging to help reduce melanoma mortality. The Memorial Sloan Kettering Cancer Center managed the project with the economic aid of philanthropic contributions (sponsors and partners). As a result, the ISIC dataset is our starting point for building the training, validation, and test set used in this contribution.

ISIC consists of several image datasets associated with well-founded clinical diagnoses: it contains more than 150.000 dermoscopy images, 7.000 of which are publicly available. Each image is associated with metadata that includes information on the image's status (benign or malignant), approximate location on the body, and demographic factors such as age and gender. ISIC is in development from 2016 to 2020. In addition, the ISIC challenge, an annual competition involving the scientific community to improve dermatologic diagnostic accuracy [36], uses ISIC to compare multiple approaches to skin disease detection.

4.3.2 Pre-processing

An ISIC refined dataset (Refined ISIC) composed of 500 RGB images (250 melanoma images and 250 benign nevi) was defined for this work. As the first step, the entire ISIC dataset was pre-processed by executing a contrast enhancement for coloured images routine (rgbCER) because many images, such as medical images, suffer from poor contrast. Therefore, enhancing such contrast of images is necessary before further pre-processing or analysis can be conducted[71]. The technique of enhancing the perceptibility of an image so that the output image is better than the input image is known as image enhancement.

Figure 4.2 shows each phase of the segmentation process; it involves the creation of a mask to split the background and the foreground. In Figure 4.2, a sample image before (a) and after the rgbCER application (b) is shown. The following code for the entire ISIC dataset was used: it executed contrast enhancement in MATLAB. Please note that the variable data in the code refer to as RGB images.

```
s_lab = rgb2lab(data);
max_luminosity = 100;
L = s_lab(:,:,1)/max_luminosity;
shadow_ad = s_lab;
shadow_ad(:,:,1) =
adapthisteq(L)*max_luminosity;
shadow_ad =
lab2rgb(shadow_ad);
data=shadow_ad;
```

The K-means segmentation was performed to identify the background and foreground in the second pre-processing step. First, the foreground mask was extracted and applied hole-filling techniques to reduce the error made by the K-means algorithm (Figure 4.2 (c)). Then, a cells detection technique to identify residual holes to dilate them (Figure 4.2 (d)) was used. Finally, a border erosion technique was performed to make the foreground mask homogeneous (Figure 4.2 (e)). The result of this second pre-processing step is the mask that highlights the background from the foreground (Figure 4.2 (f)). The following code reported the primary step of the segmentation routine.

```
Image = data;
I = uint8(image);
numColors = 2;
L = imsegkmeans(I,numColors);
B = labeloverlay(I,L);
B = imfill(B,'holes');
I = rgb2gray(I);
[~,threshold] = edge(I,'sobel');
fudgeFactor = 0.5;
BWs = edge(I,'sobel',threshold * fudgeFactor);
se90 = strel('line',5,90);
se0 = strel('line',3,0);
BWsdil = imdilate(BWs,[se90 se0]);
BWdfill = imfill(BWsdil,'holes');
BWnobord = imclearborder(BWdfill,4);
seD = strel('diamond',2);
BWfinal = imerode(BWnobord,seD);
BWfinal = imerode(BWfinal,seD);
SEGMENTED = (image.*BWfinal);
```

Figure 4.2 (g) shows the result of the segmentation approach regarding a single ISIC image. It applies the mask obtained from the previous process Figure 4.2 (f) to the image improved by rgbCER (Fig. 4.2 (b)) to exclude the background and obtain only the information relating to the foreground. Figure 4.2 (h) highlights in blue the original image parts considered foreground by the previous routine. The pixels not captured by the blue mask are considered background.

The images given as input to the algorithm are those obtained after the pre-processing phase just described, so all the images have the part of the interest in evidence and the black background. In order to provide a quality training set to our DL approach, 500

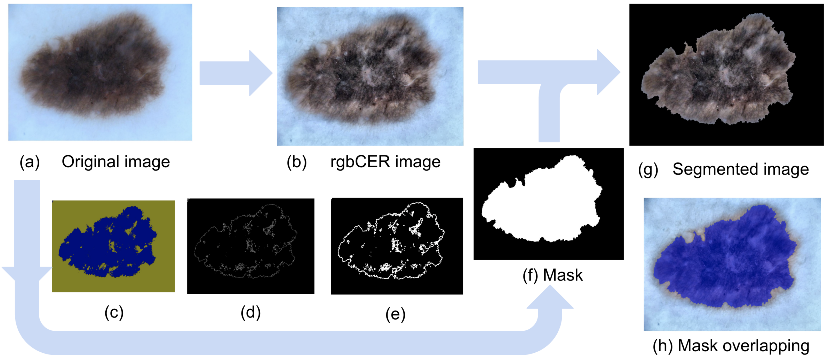


Figure 4.2: The image segmentation process.

images were manually extracted, avoiding those still containing imperfections undetected by the pre-processing step: this subset of images is called Refined ISIC in this work.

4.3.3 Training, Validation and Test sets

The complete Refined ISIC dataset was split into three subsets using the splitEachLabel function using 0.5, 0.3 and 0.2 as splitting parameters. In particular, we built three subsets named training, validation and test: training was composed of 250 images, validation of 150 images and test of 100 images. Each subset (training, validation, and test set) was built by randomly picking the images from the Refined-ISIC dataset.

Each network training session used only training and validation subsets. We used the test set to evaluate the network performances by simulating a real case scenario (in which no one of the test images was even seen by the network before).

4.3.4 Genetic Algorithms

Following the working hypothesis, GA was not used to improve hyperparameters' determination on a defined and static NN. Instead, the genetic approach was used to develop a self-assembling NN population to enhance melanoma classification. GA algorithms replicate the modes of evolution by following the Darwinian premise that the most suitable environment elements have a better chance of surviving and transmitting their features to their descendants. A population of individuals (n chromosomes) evolve from generation to generation using techniques similar to natural evolutionary processes. To represent chromosomes, we used the binary string representation; this might be a limitation because the string length limits the final size of the entity. In this experimentation, the classic GA operations were extended with another operation (merging) that allows two genomes (two entities) to concatenate together. In chromosomes, each locus (specific location on a chromosome) has two alleles (different versions of genes): 0 and 1. Therefore, it is possible to consider the chromosomes as discrete points in a solution space [49]. The evolutionary algorithms carry out heuristic exploration for new solutions to issues in which there is no complete knowledge of the search area, and they explore all of it. Then, starting with the first solution, they tweak, combine, and evolve until they find a better result. Three fundamental mechanisms are considered in evolution:

- selection: the selection indicates the process of selection of the most promising solutions capable of generating individuals who survive in the environment;
- cross over: it is a genetic recombination operator which introduces variation in the population;
- mutation: it shifts the space of solutions, resulting in the development of new information and the recovery of knowledge lost through time in the population.

As written before, another mechanism called merging was added. The merging of two chromosomes was permitted to able network structure to grow. The proposed method is described in more detail below.

4.3.5 GA definition

We report below the main definitions for the implementation of the genetic algorithms used in our work.

4.3.5.1 Population

The GA terminology is used to identify our working items. The basic components of evolutionary algorithms used in this research will be described and explained below. For example, the notation $F(t)$ indicates a composition of an object at time t.

Entity E_i is a vector $E_i = \{F_1, \ldots, F_m\}$ of m features. F_j of a generic entity E_i represents a gene of E_i. The Genome of E_i is the entire set of genes. The set $P(t) = \{E_1, \ldots, E_n\}$ is called Population at time t. For the experiment to reach a sufficiently extended network architecture, the start size of the genome was set to ten to allow at least the presence of the minimal layers required to execute a CNN. In addition, we let the genome size grow using the merge operation (unrelated to the GA fundamental) to make network architecture more complex: the merge operation concatenates two different genomes, doubling the size of the entity genome. Each gene represents one of the Matlab CNN network layers: input, dropout, batch norm, cross-chan norm, 2d-convolution, RELU, softmax, and Fully Connected. An array represents each chromosome: each cell indicates whether or not a characteristic (a feature) inside the entity exists. A feature denotes one of the layers used to build a CNN. When a feature (array cell) is active, the related layer becomes part of the network. On the other hand, we consider the feature not expressed if it is inactive and the layer does not belong to the network. Please note that a mutation may activate a non-expressed feature in the future in some evolutionary cycle. So, each feature F_j can be expressed or not by E_i and consequently, we can have silent and expressed genes. Furthermore, each chromosome does not have a predetermined length because the merging technique has been implemented and allows the joining of two chromosomes.

4.3.5.2 Fitness Function

In order to drive population evolution, we used the global population accuracy as fitness function, whose formula is reported in Equation 4.1.

$$\text{Accuracy}(ACC) = \frac{TP + TN}{TN + FP + FN + TP} \tag{4.1}$$

This fitness function represents the accuracy of all networks in an evolutionary cycle arranged in a straightforward decreasing order. The highest accuracy from each surviving entity for each evolutionary phase was used. As a result, for each generation, all entities that expose an accuracy at time t equal to or better than the previous generation at time t-*1* will survive. Also, a random 10% of the entities still survive and pass in the next evolution step, regardless of the accuracy reached at time t. The execution of the GA stops if no progress in the accuracy metric (used as a fitness function) occurs for ten consecutive evolution stages. The performance of our approach with the metrics is described in Section 2.5.

4.3.5.3 Euristic, Constraints and Limitations

Unfortunately, due to the physical restrictions of our cloud platform, the possible crossover, mutation and merging were limited to 10, 100, and 10, respectively. In order to overcome these limitations, an initial randomized population of 10.000 entities was used. The initial genomes were generated randomly. In particular, each entity's genome gene was chosen randomly from the allowed gene set and each gene parameter. In addition, the experimental environment allows running up to 100 iterations to allow the system to evolve correctly. Also, the following limitations were used:

- the initial gene of each entity E_i must be an image input (II) or one of the previously established pre-processing routines;
- if the gene g is a pre-processing routine, the gene g + *1* must be an II layer or another pre-processing layer;
- the latest gene of an entity must be a classification layer.

Due to hardware limitations, the GA could not generate graphs or cycles in the neural network structures.

4.3.5.4 Experiment execution

In order to drive the genetic evolution process, The GA engine (GAE) was implemented using C and OpenMP [17]. GAE perform all the task involved in population initialization and management. In particular, each simulation step (SS) is split into three phases: population evolution (P1), Network execution (P2) and fitness evaluation (P3).

- Phase P*1*: for each SS, the GAE performs the GA operations (mutation, crossover and merging) on the current population to obtain new NN candidates;
- Phase P*2*: GAE sends the entire population and training sets to a GRIMD cluster using a map/reduce approach: each NN is associated with a worker; each worker calls the MATLAB train function to train the NN on the training datasets. At this point, GAE will hold, waiting for the completion of the “reduce" phase that will return to GAE the survived trained networks with the corresponding accuracy.
- Phase P*3*: GAE sorts all the accuracies in descending order and selects the new candidate for the next evolution, following the rules described in the fitness section.

Figure 4.3 shows a simplified overview of the computational environment.

4.3.6 Experiments Setup

The experiment was performed with a hybrid Beowulf/Cloud Computational (GRIDC) architecture setup designed to run the `Matlab 2021` environment across multiple workstations and cloud workers. In addition, we adapted GRIMD architecture [73] that provides a map/reduce approach to distributing across the “grimd slaves" the working package composed of: training and validation sets and neural networks structure to train.

The Beowulf part of the GRIDC was composed of three high-performance workstations equipped with NVIDIA GPU, high RAM available and multicore capabilities running Windows 10:

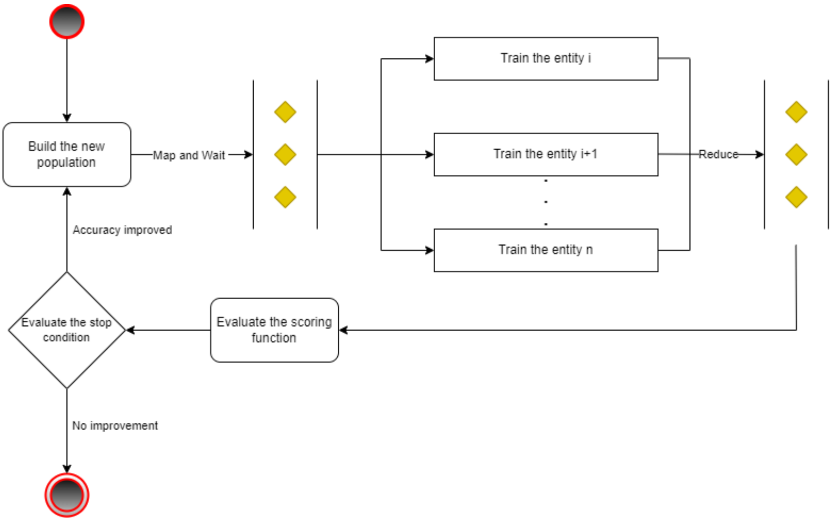


Figure 4.3: An overview of the experiment environment architecture.

the SoftMatterLab (University of Salerno – DIFARMA) provided pro-bono the Beowulf part of GRIDC.

The cloud part was composed of three Amazon AWS c5d.metal instances running Centos. However, we shut down the c5d.metal cluster during the experimentation as soon as the number of NN became tractable only via our HPC system. The cloud part was provided pro-bono by Softmining SRL.

Table 4.2: Work Environment

GRIDC	#Core	RAM	GPU0	GPU1
Beu	64	250GB	Quadro P400	RTX5000
Beu	20	64GB	Intel UHD G 770	GeForce RTX 3060 Ti
Beu	12	16GB	Intel UHD G 630	Quadro P2000
Cloud	96	192GB	-	-
Cloud	96	192GB	-	-
Cloud	96	192GB	-	-

4.3.7 Results and Discussion

This section reports and analyzes the preliminary results of our experiments.

Immediately after the initialization phase, at the second iteration, the population shrank by a ratio of about 1:3. In particular because the NN structure (NN structure that causes MATLAB train function to crash) was not compatible or because the NN structure was valid but requires more RAM than the available RAM on workers: this causes high death ratio. In particular, the second observation seems strictly related to the initial choice to avoid pooling layers to be part of the genome. These facts strongly suggest that an improvement in the random initialization approach (better heuristics) is needed. Figure 4.4 reports the Trend of the death ratio across the eleven iterations.

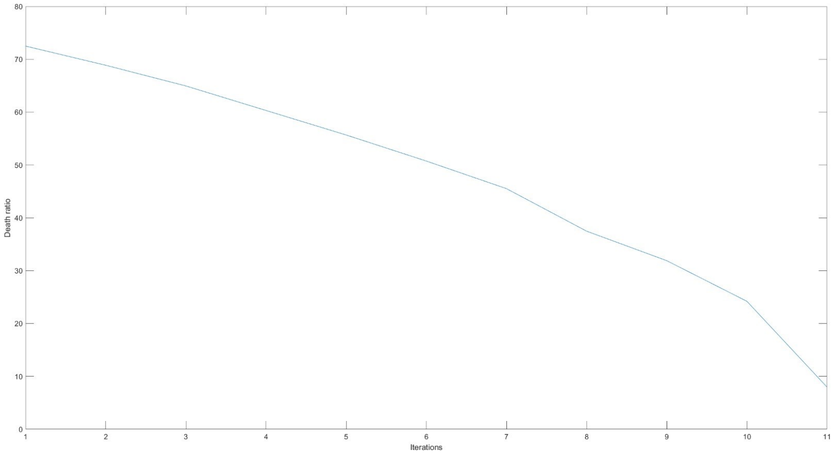
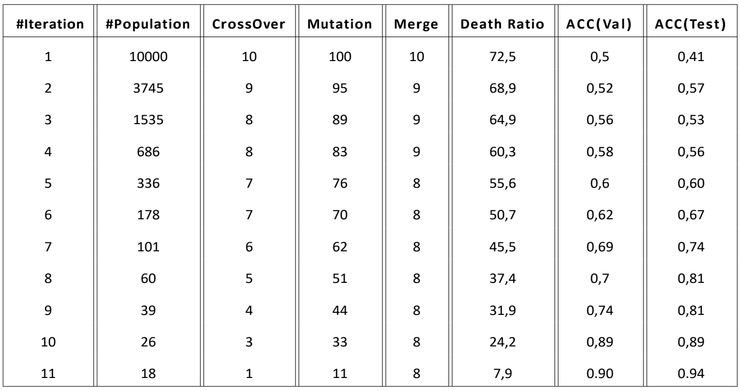


Figure 4.4: Trend of death ratio over the 11 iterations.

However, despite the high death ratio, the execution of the GA led to an improvement in the accuracy metric in every iteration by selecting entities with high-quality criteria as the best option over less desirable alternatives. However, many NN populations tended to overfit: preliminary observations suggest many dropout layers are needed. Also, it is possible to see that the GA algorithm reached a stable population (a plateau regarding accuracy) after 11 evolutionary cycles without reaching the limit of 100 iterations; at the 11th iteration, it appears that the loss of diversity of the NN population prevents the exploration of novel solutions because the GA reached a local or global optimum. This behaviour led to drastic computation time and cost-saving because the number of “networks to train" rapidly became tractable without cloud computational power. Furthermore, it was possible to disable the "c5d.metal" (Table 4.2) instances after the fourth evolutionary cycle: the residual evolutionary steps needed only up to three days to complete.

The high mortality ratio observed may be due to the randomness of the initial population, where the networks die either because the layers are incompatible with each other or because the structure is incompatible with the working environment.

Table 4.3: Evolution during the iterations



#Iteration	#Population	CrossOver	Mutation	Merge	Death Ratio	ACC(Val)	ACC(Test)
1	10000	10	100	10	72,5	0,5	0,41
2	3745	9	95	9	68,9	0,52	0,57
3	1535	8	89	9	64,9	0,56	0,53
4	686	8	83	9	60,3	0,58	0,56
5	336	7	76	8	55,6	0,6	0,60
6	178	7	70	8	50,7	0,62	0,67
7	101	6	62	8	45,5	0,69	0,74
8	60	5	51	8	37,4	0,7	0,81
9	39	4	44	8	31,9	0,74	0,81
10	26	3	33	8	24,2	0,89	0,89
11	18	1	11	8	7,9	0.90	0.94

Table 4.3 shows the Evolution during the 11 iterations.

After the stop of GA routines, 19 NN (19NN set) exposed a validation accuracy less or equal to 0.90, with a final validation loss of 0.2047: due to the high validation accuracy, all these 19 networks were considered equivalent. The Figure 4.5 shows the training plot of one network picked from the 19NN set. The accuracy of this network on the test set was 0.94.

Also, Figure 4.6 shows the confusion matrix regarding network performance on the test set. The crossover and the mutation operations tend to decrease following the population size and accuracy.

The preliminary results reported in this work strongly suggest that the GA approach can enable the design of the structure of a neural network driven by the problem to solve, avoiding human

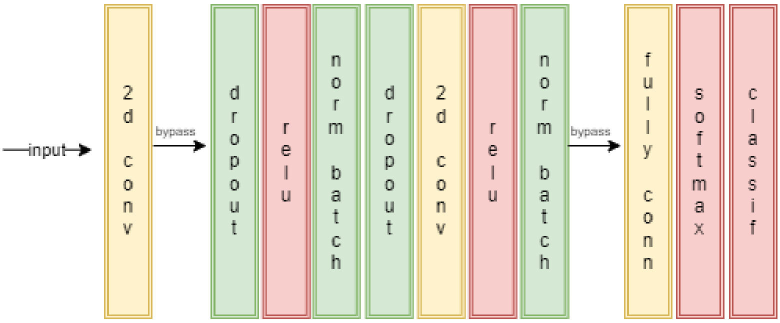


Figure 4.5: Structure of the best network in the last iteration.

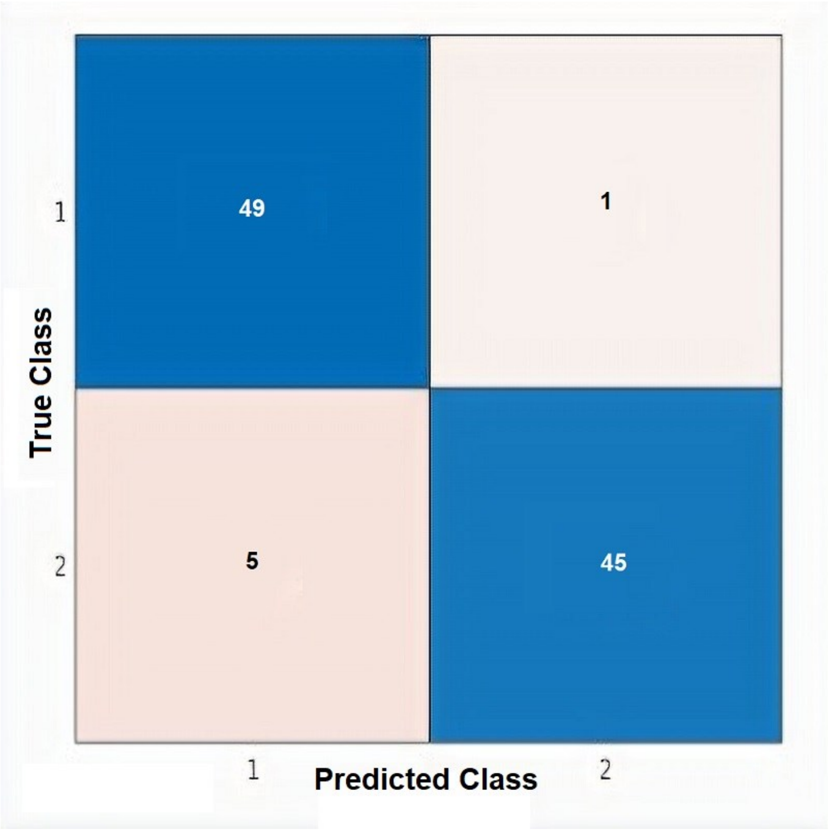


Figure 4.6: Confusion matrix of the best NN in the 11th iteration.

interaction. Table 4.4 shows the best results for each iteration until convergence is reached (up to the 11th iteration). The last iteration reports excellent performances, such as 0.97 and 0.98 of specificity and precision, which show the potential to outperform the method proposed in the literature. However, performing more experimentation, particularly performance analysis, is mandatory, using more extensive training and test sets.

Table 4.4: Experiments results

#Iteration	TP	FP	FN	TN	ACC	SEN	SPE	PRE
1	36	14	42	8	0,44	0,46	0,36	0,72
2	50	0	43	7	0,57	0,53	1	1
3	50	0	47	3	0,53	0,51	1	1
4	41	9	35	15	0,56	0,53	0,62	0,82
5	10	40	0	50	0,60	1	0,55	0,2
6	34	16	17	33	0,67	0,66	0,67	0,68
7	50	0	26	24	0,74	0,65	1	1
8	44	6	13	37	0,81	0,77	0,86	0,88
9	50	0	19	31	0,81	0,72	1	1
10	43	3	8	42	0,89	0,85	0,93	0,94
11	49	1	5	45	0,94	0,90	0,97	0,98

4.3.8 Comparison with the literature

Table 4.5 reports the performances of multiple deep-learning techniques performed on ISIC datasets. For our proposal, a balanced and small refined subset of images of ISIC, composed of 500 images, was used to provide quality images to the networks. On this refined dataset, the approach, which involves building the network guided by the use of genetic algorithms, achieves maximum values of ACC of 94%, SEN of 90%, SPE of 97%, and PRE of 98% at the eleventh iteration. This data was used for a preliminary comparison with other approaches.

Alizadeh et al [3] proposed an ensemble method to detect melanoma. Their method consists of two models of CNNs and two texture features, local binary pattern and Haralick features. During ISIC 2016 and ISIC 2019, the authors evaluated the

Table 4.5: Reference literature

Reference Paper	Dataset	Methods	ACC	SEN	SPE	PRE
Alizadeh et al [3]	ISIC 2016	CNN + feature extraction	85.2%	52%	93.40%	66%
Alizadeh et al [3]	ISIC 2019	CNN + feature extraction	96.7%	96.3%	97.1%	95.1%
Kaur et al [50]	ISIC 2016	LCNet	81.41%	81.3%	80.83%	81.88%
Kaur et al [50]	ISIC 2017	LCNet	88.23%	87.86%	88.86%	78.55%
Kaur et al [50]	ISIC 2020	LCNet	90.42%	90.39%	90.39%	90.48%
Our approach	**ISIC 2020**	**GA design**	**94%**	**90%**	**97%**	**98%**

method. The first model has nine layers and employs many batch normalization layers to speed up classification and prevent the problem of overfitting. The second model used a pre-trained VGG-19. An ensemble approach is used to integrate these two models for the classification task.

In 2022, Kaur et al. [50] proposed a deep convolutional neural network, called LCNet, to classify malignant versus benign melanoma images. The deep network is composed of eleven blocks.

4.3.9 Conclusion

This contribution proposes using genetic algorithms to build CNNs to address the melanoma classification problem, one of the most dangerous skin cancers. The initial hypothesis claims that it is possible to interpret a network's development as a system's evolution over time. In the GA context, the entire system adapts, modifying its configuration in response to the dynamic of its interactions with the environment, like reinforcement learning. Finally, the system leads to selecting optimal solutions (local or global) to achieve the goals of the considered task.

The initial generation of NN is stochastic. Consequently, initially, a remarkably low accuracy is observed, while with the advancement of the experiment, there is an attenuation of the error consequent to a better fitness level of the NN population. A set of equivalent NN with high classification performance was available in the last evolutionary iteration.

According to the preliminary results, allowing GA to assist in designing a NN structure could yield results comparable to traditional NN design (by humans) methods. Furthermore, the proposed approach must be expanded and evaluated on larger or additional melanoma datasets (e.g., clinical or histological). The future goal aims to extend the training set images to improve the understanding of the genetic algorithm in constructing the network as the starting dataset increases.

Also, evaluating a new heuristic could reduce the high death ratio and the tendency to overfit and permit fast convergence to a local or global optimum. In future research, using additional criteria to define the initialization of the algorithm and the use of a more targeted population to achieve the desired result might be investigated. Also, extending the permitted layer to be part of the entity genome might be considered to allow GA to explore more solutions. Finally, We will also plan to improve the selection procedures of individuals to be used for subsequent generations through the fitness function.

5 A CLOUD APPROACH FOR MELANOMA DETECTION

5.1 background and considered issues

Early detection of melanoma is crucial for improving life expectancies, especially in people who are at high risk of developing the disease. Due to the many visual similarities between melanoma and non-melanoma, early detection is particularly difficult [95]. Nowdays, there are a plethora of proposals for a computer-aided system for dermatologists, especially with the use of CNNs [74]. In this work, we proposed a more adaptable system design that can deal with modifications to the training datasets. To provide a Melanoma Detection service based on clinical and dermoscopic images, we suggested the development and application of a hybrid architecture based on cloud, fog, and edge computing. This architecture must simultaneously deal with the volume of data that needs to be evaluated by reducing the running time of the continuous retrain. Specifically, the proposed hybrid architecture is composed by:

- Cloud layer: Where data storage and high-performance computing operations are performed. Also finished in the cloud are the stages for validation and testing. Whenever a new classifier becomes available, the Cloud layer will send a new network to the edges;

- Fog layer: Data from the Edge layer is received by network-distributed server systems, which pre-process, filter, and post the data to the Cloud. Mid-weight computational techniques can also be used at this level. Finally, the Fog layer may employ the same methodology as the Edge layer for devices with insufficient computing power;

- Edge layer: It consists of every intelligent IoT architectural device, or Edge Device. At this level, the Edge Device processes the data. The most recent classifier is run in the Edge in order to examine the images.

In spite of the different techniques used for the classification of melanoma, the following work has focused on two open problems:

1. the transfer learning reliability evaluation;
2. the impact of Impact of the three-layers architecture.

Transfer learning is a technique that speeds up training durations when applying previously learned information to a related problem by allowing the use of pre-trained networks. This transfer is based on the assumption that the target data and the original data are in the same feature space and have the same distribution [69]. Before network training, the three datasets (training, validation, and test) are frequently fixed. A slight change in the subsets may have an effect on prediction accuracy, a fact that may go unnoticed. This indicates that Transfer Learning is still not reliable. Consequently, related to the first open issue, we want to demonstrate how changes to the structure of a dataset could result in a reduction in the system's overall performance.

For the second open issue, we measured the performance of the proposed architecture. To demonstrate that a distributed and cooperative system is required to deploy a melanoma classifier robust against Transfer Learning difficulties, we specifically build the architecture to allow automatic classifier retraining and deployment. Particularly, since data structure changes, a huge number of iterations are required to get the optimal classifier. In this configuration, the Fog layer, which stores and sends each image to the Cloud, allows the user to query the system and participate in dataset and model changes. In this case, the data scientists just need to classify additional images (e.g., melanoma/non-melanoma) as they are acquired. In Figure 5.1, we depicted the proposed three-layers architecture.

5.2 proposed method

We simulated a three-layer architecture, with the cloud layer serving as the training and retraining layer. The GRIMD framework was used to build this system, allowing us to distribute each iteration among numerous instances [73]. We have first set up the GRIMD instances on Amazon AWS. The training, retraining, validation, test, and performance comparison processes were then transferred to the cloud layer. The fundamental element is that a new model is only deployed into Fog when its accuracy outperforms that of the one that came before it. The Layer Agents, which we developed as a straightforward CROND instance, con-trol the synchronization between each layer. Finally, the trained models and web server were also stored in the Fog layer together with the classification and prediction procedures. Every end user in this scenario interacts with the Fog layer via an app. Therefore, we may conclude that using a distributed design could offer the end user various advantages by offering: the gathering and assembling of data on the network to aid in the early detection of melanoma, enhancing image databases with new information; processing crucial data locally at the network's edge with local data storage leads to decreased data processing latency, real-time response, lower bandwidth, and faster data access; widespread distribution of resources and computing services, made possible by a large number of mobile Fog nodes. The issue of delivering images to a central data server or Cloud service for processing is specifically addressed by this architecture. Decentralizing them also improves the capacity and, thus, the calculating times.

Figure 5.1 displays how the suggested hybrid design works overall. Data buckets are kept up-to-date, and system training is done in the cloud. In the Fog area, where services are executed, the orchestrator is in charge of distributing the optimal services following each formation. The Edge area is where local calculations are carried out on IoMT devices (such as smartphones). An early examination of the loaded data is carried out by the software program HiC-Otsu, which is part of the Fog system on the IoMT device. To enhance the efficiency of the system, the QoS moderator annotates content. Although the average user takes use of the output of the services, he also adds to the system's knowledge base by loading data.

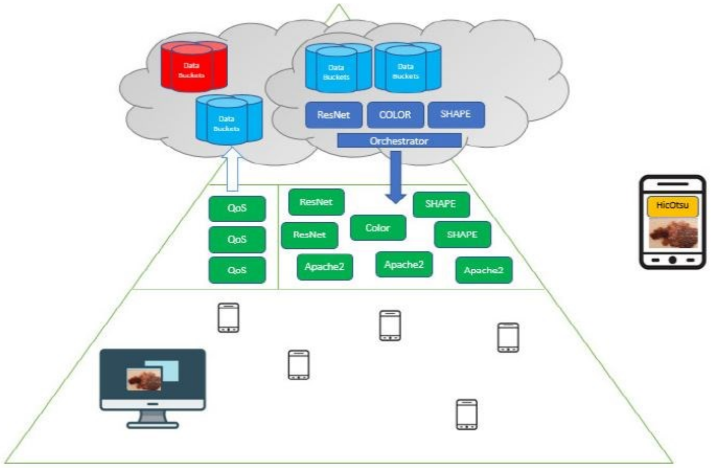


Figure 5.1: General operation of the three layers architecture for melanoma detection.

5.3 used networks

We compared the principle neural networks: AlexNet, GoogleNet and Google InceptionV3. AlexNet is an eight-layer convolutional neural network; the first five layers were convolutional, some of them were followed by max-pooling layers, and the final three layers were completely connected. It made use of the non-saturating ReLU activation function, which outperformed tanh and sigmoid in terms of training performance [40].

GoogleNet is a convolutional neural network with 27 layers that is made up of around 100 different building blocks, including convolutions, average pooling, maximum pooling, and contacts. This network is based on the core Inception design, which debuted in 2015 and is a computationally effective network even with constrained computer resources. Through Google Cloud Platforms, GoogleNet executions on Cloud TPU are accessible [89].

The development of the Inception Architecture is a key component of Google InceptionV3 [20]. On the ImageNet dataset, it has been demonstrated to achieve higher than 75% accuracy. It is a widely used image recognition model. Convolutions, average pooling, max pooling, concerts, dropouts, and fully linked layers are some of the symmetric and asymmetric building pieces that make up Google InceptionV3. The model makes considerable use of batchnorm and applies it to activation inputs. Using Softmax, loss is calculated [89].

5.3.1 MED-NODE dataset and Pre-Processing

The used dataset is MED-NODE computer-assisted system for melanoma diagnosis (here named MED-NODE), composed by 170 clinical images (70 melanoma and 100 nevi images) from the digital image archive of the Department of Dermatology of the University Medical Center Groningen (UMCG) [33]. Dermatolo-gists have checked each image's accuracy before labeling it. The images are from various Caucasian patients and have previously undergone pre-processing and anonymization. Hair has already been removed with the Dullrazor program [58].

Four additional datasets (MDS) have been developed as a result of our assumption that Transfer Learning is not realiable, as will be discussed below:

- **MD1**, which contains MED-NODE original images;
- **MD2**, which contains MED-NODE images segmented with the Otsu method;
- **MD3**, which contains MED-NODE images and augmented images without segmentation;
- **MD4**, contains MED-NODE images and augmented images segmented with the Otsu method.

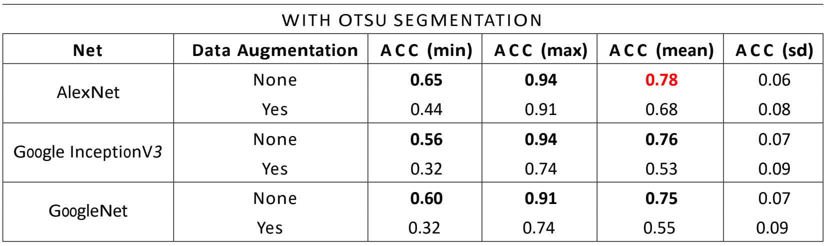
The fundamental premise is that, when four datasets are generated, the source and destination domain data may differ in terms of the marginal distribution, but that the reference labels will always be the same. For each dataset D in MDS, we repeated the training phase 700 times to simulate continual retraining. The dataset was divided using the following ratios for each iteration: 0.5 for the training set, 0.3 for the validation set, and 0.2 for the test set.

Due to the large amount of previous research in this area, the work did not concentrate on the features extraction stage (which includes segmentation, boundary analysis, and other characteristics). In the segmentation/pre-processing step, the most basic Otsu segmentation and Gaussian filter are employed. The Otsu approach can reduce intra-class variation [67]. An application of the imgaussfilt function with a dynamic sigma value between 1 and 7 was utilized to remove noise before each network training [8]. We performed 100 training iterations for each sigma value. Moreover, since the dataset only contains 170 total images (70 malignant and 100 benign), we choose the data augmentation technique to introduce more variants by manipulating the images artificially [85]. We opted for scale (in the range 1-10), translation on X and Y axes and rotation (all in the range -180, 180).

5.4 results and discussion

5.4.1 Results for the transfer learning reliability evaluation

In Table 5.1 and Table 5.2, we present the findings from the analysis of the MED-NODE dataset with and without Otsu segmentation.



WITH OTSU SEGMENTATION

Net	Data Augmentation	ACC (min)	ACC (max)	ACC (mean)	ACC (sd)
AlexNet	None	**0.65**	**0.94**	**0.78**	0.06
	Yes	0.44	0.91	0.68	0.08
Google InceptionV3	None	**0.56**	**0.94**	**0.76**	0.07
	Yes	0.32	0.74	0.53	0.09
GoogleNet	None	**0.60**	**0.91**	**0.75**	0.07
	Yes	0.32	0.74	0.55	0.09

Table 5.1: Performance on MED-NODE dataset for ACCs with Otsu segmentation and with and without data augmentation

The highest values that the networks achieved in the computations of the average, maximum, minimum, and standard deviation values of the ACC are shown in bold. The AlexNet network achieves the best outcome for the average ACC both with and without using data augmentation or segmentation (highlighted in red). According to the equation in Equation 5.1, we also used the standard deviation (SD) to estimate how much the accuracy measures varied between the networks.

WITHOUT OTSU SEGMENTATION

Net	Data Augmentation	ACC (min)	ACC (max)	ACC (mean)	ACC (sd)
AlexNet	None	0.68	**1**	**0.89**	0.05
	Yes	**0.76**	0.97	0.87	0.05
Google InceptionV*3*	None	**0.56**	**0.94**	**0.74**	0.07
	Yes	0.32	0.71	0.55	0.07
GoogleNet	None	**0.65**	**0.94**	**0.80**	0.06
	Yes	0.30	0.76	0.55	0.09

Table 5.2: Performance on MED-NODE dataset for ACCs without Otsu segmentation and with and without data augmentation

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (5.1)$$

where n is the size of the dataset and $\overline{x}$ is $\frac{1}{n}\sum_{i=i}^{n} x_i$ the arithmetic mean of x. In Figures 5.2(a)-5.2(c) are reported the SD values for all three used networks.

In Table 5.3 and Table 5.4, in addition, we reported the values of sensitivity (TPR), specificity (TNR), precision (PPV), false discovery rate (FDR), false-negative rate (FNR) and false-positive rate (FPR) for the three networks.

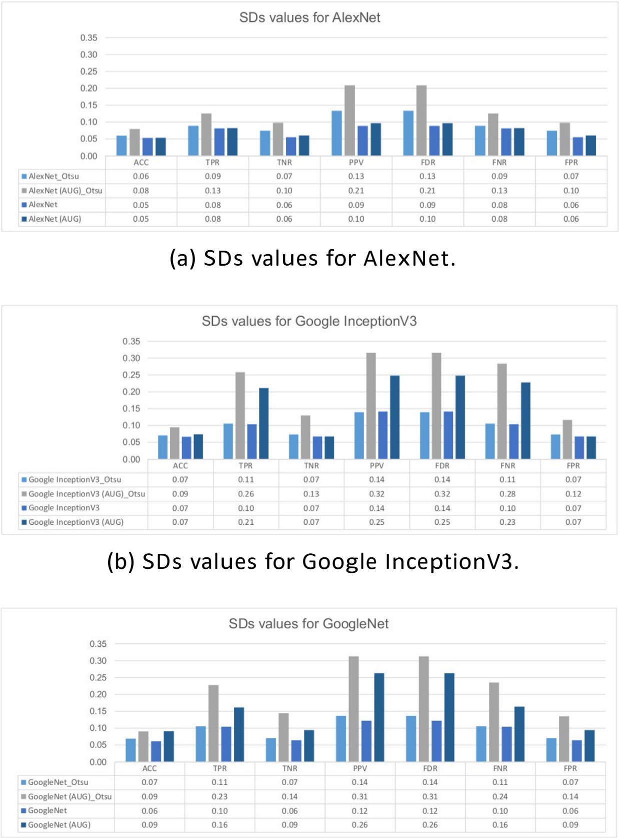
WITH OTSU SEGMENTATION

Net	Data Augmentation	TPR (%)	TNR (%)	PPV (%)	FDR (%)	FNR (%)	FPR (%)
AlexNet	None	75	82	73	27	25	18
	Yes	63	76	65	35	37	24
Google InceptionV*3*	None	74	79	68	32	26	21
	Yes	38	57	38	62	47	42
GoogleNet	None	72	78	67	33	28	22
	Yes	44	62	40	60	51	37

Table 5.3: Performance on MED-NODE with Otsu segmentation and with and without data augmentation

5.4.2 Results for the impact of the three-layers architecture

For each dataset, we examined the behaviors of the networks to assess their performance. The calculations in this second experiment are identical to those in the first, but they were scaled up



(a) SDs values for AlexNet.

(b) SDs values for Google InceptionV3.

(c) SDs values calculated for GoogleNet.

Figure 5.2: Several SDs values computed for all networks.

WITHOUT OTSU SEGMENTATION

Net	Data Augmentation	TPR (%)	TNR (%)	PPV (%)	FDR (%)	FNR (%)	FPR (%)
AlexNet	None	87	90	86	15	13	10
	Yes	84	91	87	14	16	9
Google InceptionV3	None	73	76	62	38	27	24
	Yes	39	60	29	71	57	40
GoogleNet	None	79	82	72	28	21	18
	Yes	45	63	48	52	54	37

Table 5.4: Performance on MED-NODE without Otsu segmentation and with and without data augmentation

to 128 GB and several GPUs by employing Ec2 instances of types t2 (micro-instances: t2.micro) and m5 (balanced computation instances: m5a.2xlarge), up to type c6 (optimized computation instances: c6g.16xlarge).

The results, shown in Table 5.5, indicate that GoogleNet is the most reliable network, with a mean decline in prediction accuracy of -19.60%.

Net	Measure	MD1	MD2	MD3	MD4	Mean Drop
AlexNet	Best	0.97	0.91	0.97	0.89	
	Average	0.81	0.72	0.81	0.73	
	Drop	-19.75	-26.38	-19.75	-21.91	-21.95
Google InceptionV3	Best	0.91	0.88	0.90	0.89	
	Average	0.75	0.72	0.75	0.74	
	Drop	-21.33	-22.22	-20.0	-20.27	-20.96
GoogleNet	Best	0.94	0.93	0.91	0.89	
	Average	0.81	0.77	0.75	0.74	
	Drop	-16.04	-20.77	-21.33	-20.27	**-19.60**

Table 5.5: Performance drop after 100 training steps (related to Training and Validation steps)

Environment	GoogleNet	Google InceptionV3	AlexNet
Single	82710	115200	19724
GRIMD(t2)	55140	94348	13327
GRIMD(m5)	20677	37105	6872
GRIMD(c6)	7519	17710	3171

Table 5.6: Clock time (in seconds) measured for both the experiments

Table 5.6 displays the clock times, in seconds, for both experiments. We choose the clock time because we intended to calculate the time saved by data scientists under two conditions:

1. mimicking the standard scenario in experiment 1, which requires for laboratory effort to undertake training, validation, and deployment as well as update the datasets;

2. In the case of experiment 2, a system is offered with the exception of image annotation, which is in any case left to qualified dermatologists and handles all other aspects.

The amount of time and effort required to keep a classifier operating at its best was gathered. We invested up to 82000 seconds in each retraining in order to achieve a good result for the MED-NODE datasets (which only contain 170 images).

5.5 conclusion

The results of this research point to the possibility that, despite the great performance noted in the literature, the widely used Transfer Learning approach may not be reliable. These results agree with what recently happened in [34], when, due to the addition of new categories and images, the ISIC 2018 winning algorithm performance decreased from 88.5% to 63.6% in the ISIC 2019. In this work, we performed two experiments: the first is focused on the evaluation of the TL approach; the second is based on the advantages provided by the use of an architecture based on Cloud, Fog and Edge layers. Results from the first experiment, in particular, demonstrate how a classifier's performance could deteriorate by even little modifications. Additionally, according to our conclusions, AlexNet is the most reliable network in terms of the Transfer Learning problem. Continuous retraining is necessary to prevent performance loss since many training iterations are necessary to get the optimal classifier. Based on the results for the second experiment, we were able to save up to 76% of computational time by carrying out the continuous retraining process required to keep the robustness of the classifier. Moreover, our results demonstrated that CNN networks reported better performance without segmentation. This finding might imply that training should take into account information contained in the skin around lesions.

x

6 CONCLUSIONS

In this Chapter, a summary of the research contributions is reported in Section 6.1. Also, in Section 6.2, some suggestions for future work are discussed.

6.1 summary

In this thesis, four open issues in melanoma detection and classification fields are addressed: the first concerns the change in POV from the maximization of the accuracy to the minimization of the life-threatening scenarios, in particular by minimizing the false-negative rate event at the expense of the global accuracy; the second concerns the utilization of clinical images, alone or together with other kinds of images, to speed up the early diagnosis, in particular regarding the fact that nowadays there is a massive amount of mobile devices able to take the picture at high resolution, allowing the creating of extensive training dataset without the need of specific instrument like the dermoscopic; the third is related to the research and design of new CNN models able to minimize the FNR using genetic algorithms; the latest concerns the design of architecture able to address the intra-class dissimilarities/ extra class similarities allowing a cloud-fog-edge system to perform continuous and robust retraining.

The final contribution of this work is the proposal of a guideline for designing and implementing a robust system that can withstand performance decreases that may occur when the training data changes, as reported in the chapter 5.

6.1.1 Changing the POV regarding CAD performances

As discussed before, Melanoma is a type of skin cancer considered one of the world's most dangerous and deadly tumour forms that start from the melanocytes, and in its early stages, it may be mistaken for a regular naevus. Although it accounts for a small portion of skin cancers, it is the leading cause of death among those diagnosed: this highlights the significance of early detection of Melanoma, particularly for individuals who are at a higher risk of developing the disease, as this increases the chance of a successful cure. Early detection also plays a critical role in the effectiveness of first-line treatment for this type of cancer. In order to increase the early detection effectiveness, in recent years, there has been a significant and rapidly growing increase in the data available on melanoma images and cure outcomes. This data was used for correlation studies to build automatic classificators and better understand melanoma disease evolution.

These tools showed the potential to become essential in producing more accurate diagnoses, developing new treatments and gaining new insights and knowledge in the next future, but despite the high performance reported for these tools, particularly the high accuracy, nowadays, the last word on the diagnosis remains to the dermatologist because, in the case of a malignant suspect, a biopsy is needed. After all, it is impossible to fully trust the results obtained from these tools, in particular, due to the intra-class/extra-class issues described in chapter 5. that showed how the accuracy of these tools could drop dramatically.

The most hazardous situation during melanoma assessment is the occurrence of a physician's false positive (FP) or false negative (FN) diagnosis: this is true also for automatic tools. In the case of a false positive, the patient may experience added stress and anxiety due to the fear of Melanoma and the need for a biopsy, only to find out later that they do not have cancer. On the other hand, in the false negative scenario, the patient may already have Melanoma in its early stage and could have been fully cured if correctly diagnosed. In a false positive case, if the physician fully trusts the automatic system, he may have deemed the skin lesion benign in this case, thus not requiring further examination. While the stress and fear of a false positive diagnosis can be survived, the progression of Melanoma in the false negative scenario may lead to a grim outcome: this highlights the importance of research efforts aimed at developing automatic tools that prioritize reducing false adverse events, even if it means an increase in false positive diagnoses.

6.1.2 The potential contribute of the clinical images

The second and third contributions provide a new way of classifying medical images, using clinical images instead of dermatoscopic and histological images. Advances in technology have allowed us to process and analyze medical images using mathematical algorithms to uncover information and gain a deeper understanding of pathological and physiological processes that cannot be detected through visual analysis alone. Additionally, using clinical images in classification models can lead to the broader adoption of CAD systems, as there is no need for specialized hardware such as dermatoscopic cameras due to the high-resolution cameras commonly found on mobile devices. This second contribution delves into the performance of the main available CNN architectures in minimizing the FNR when utilizing clinical images.

The results were achieved by combining the main capabilities of Genetic Algorithms (GA) with Convolutional Neural Networks (CNN) to address the melanoma detection problem (GACNN). The outcome was achieved by combining the strengths of Genetic Algorithms (GA) and Convolutional Neural Networks (CNN) to tackle the problem of melanoma detection (GACNN): These algorithms enable parallel processing and the attainment of near-excellent results in reasonable time frames (refer to Chapter **??**). However, although this strategy leads to an acceptable solution, further clarification of the starting parameters of the algorithms and the associated genetic functions (such as selection, crossover, and mutation) is still needed.

The results from these experiments indicate that allowing GA to construct the CNN structure can enhance melanoma classification by enabling the network design to evolve without human intervention. Furthermore, the proposed approach can be expanded to other future melanoma datasets, such as clinical, dermatoscopic, or histological images, using other innovative evolutionary optimization algorithms.

6.1.3 Addressing the intra-class/extra-class issue with the continuous retraining

In order to address the intra-class/extra-class issues (ICEC), the continuous retraining approach was proposed. However, this approach needs high computational power and high storage space. In order to allow the implementation of continuous retraining, a scalable three-tier architecture (Cloud, Fog, and Edge) is suggested. In particular, the system proposed aimed to address the issues of storage, training/retraining, and distribution of models for Melanoma classification. The primary idea is to provide an architecture where common users can effortlessly create and insert new classification models without altering the three-tier architecture.

The proposed architecture was used to study the robustness of three deep neural networks (AlexNet, GoogleNet, and Inception V3) against the ICEC issues: AlexNet was the most stable network, while all the CNN tested showed improvement in their average accuracy without the use of segmentation or data augmentation. These findings encourage the utilization of continuous retraining to reduce false positives and increase sensitivity.

The main contributions and accomplishments of this thesis can be summarized as follows:

- a profound overview of the Melanoma diseases that lead to the need to design CAD focusing on FNR minimization instead of maximization of the accuracy;
- results that show the potential of the utilization of the clinical images for the early diagnosis of Melanoma;
- results that show the potential of the utilization of the genetic algorithm to design automatic way CNN structures oriented to minimize the FNR;
- results that show the main CNN architectures performances using clinical images instead of dermatoscopic images;
- results that suggest that continuous retraining may address the ICEC issues;

- a proposal of three-tier architectures that enable the execution of continuous retraining;

6.2 future works

Deep learning algorithms have been able to achieve impressive results in many fields. However, with the increase in Big Data, algorithms and platforms are needed to develop further to keep up with the growth. One area of focus for improving the accuracy of melanoma classification systems based on neural networks is to include a broader perspective that considers all relevant data, including anamnestic data related to the patient and their family history and additional clinical features that can be extracted from images. In the future, the team intends to investigate the correlations between different data sources and explore the use of transfer learning methods to consider the heterogeneity of multi-domain text sources.

BIBLIOGRAPHY

[1] Naheed R Abbasi, Helen M Shaw, Darrell S Rigel, Robert J Friedman, William H McCarthy, Iman Osman, Alfred W Kopf, and David Polsky. "Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria." In: Jama 292.22 (2004), pp. 2771–2776.

[2] Wiem Abbes, Dorra Sellami, Stella Marc-Zwecker, and Cecilia Zanni-Merk. "Fuzzy decision ontology for melanoma diagnosis using KNN classifier." In: Multimedia Tools and Applications 80 (2021), pp. 25517–25538.

[3] Seyed Mohammad Alizadeh and Ali Mahloojifar. "Automatic skin cancer detection in dermoscopy images by combining convolutional neural networks and texture features." In: International Journal of Imaging Systems and Technology 31.2 (2021), pp. 695–707.

[4] Alper Arik, Mesut Gölcük, and Elif Mine Karslıgil. "Deep learning based skin cancer diagnosis." In: *2017 25*th Signal Processing and Communications Applications Conference (SIU). IEEE. 2017, pp. 1–4.

[5] International Agency for Research on Cancer Working Group on Artificial Ultraviolet (UV) Light and Skin Cancer. "The association of use of sunbeds with cutaneous malignant melanoma and other skin cancers: a systematic review." In: International Journal of Cancer 120.5 (2007), pp. 1116–1122.

[6] Alessia Auriemma Citarella, Luigi Di Biasi, Michele Risi, and Genoveffa Tortora. "SNARER: new molecular descriptors for SNARE proteins classification." In: BMC bioinformatics 23.1 (2022), pp. 1–20.

[7] Wei Ba, Huan Wu, Wei W Chen, Shu H Wang, Zi Y Zhang, Xuan J Wei, Wen J Wang, Lei Yang, Dong M Zhou, Yi X Zhuang, et al. "Convolutional neural network assistance significantly improves dermatologists' diagnosis of cuta-

neous tumours using clinical images." In: European Journal of Cancer 169 (2022), pp. 156–165.

[8] Mitra Basu. "Gaussian-based edge-detection methods-a survey." In: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 32.3 (2002), pp. 252–260.

[9] Abi Berger. "How does it work?: Positron emission tomography." In: BMJ: British Medical Journal 326.7404 (2003), p. 1449.

[10] Prachi Bhave, Lalit Pallan, Georgina V Long, Alexander M Menzies, Victoria Atkinson, Justine V Cohen, Ryan J Sullivan, Vanna Chiarion-Sileni, Marta Nyakas, Katharina Kahler, et al. "Melanoma recurrence patterns and management after adjuvant targeted therapy: a multicentre analysis." In: British journal of cancer 124.3 (2021), pp. 574–580.

[11] National Institute of Biomedical Imaging and Bioengineering. Computed Tomography (CT). Last accessed June 2022. 2022. url: https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct.

[12] Salah Bouktif, Ali Fiaz, Ali Ouni, and Mohamed Adel Serhani. "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches." In: Energies 11.7 (2018), p. 1636.

[13] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Stefan Fröhling, et al. "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task." In: European Journal of Cancer 111 (2019), pp. 148–154.

[14] Alessia Auriemma Citarella, Fabiola De Marco, Luigi Di Biasi, Michele Risi, and Genoveffa Tortora. "PADD: Dynamic Distance-Graph based on Similarity Measures for GO Terms Visualization of Alzheimer and Parkinson diseases." In: J. Vis. Lang. Comput. 2021.1 (2021), pp. 19–28.

[15] Alessia Auriemma Citarella, Lorenzo Porcelli, Luigi Di Biasi, Michele Risi, and Genoveffa Tortora. “Reconstruction and Visualization of Protein Structures by exploiting Bidirectional Neural Networks and Discrete Classes.” In: *2021 25*th International Conference Information Visualisation (IV). IEEE. 2021, pp. 285–290.

[16] Luigi D’Arco, Huiru Zheng, and Haiying Wang. “SenseBot: A Wearable Sensor Enabled Robotic System to Support Health and Well-Being.” In: CERC. 2020, pp. 30–45.

[17] Leonardo Dagum and Ramesh Menon. “OpenMP: an industry standard API for shared-memory programming.” In: IEEE computational science and engineering 5.1 (1998), pp. 46–55.

[18] Lauren E Davis, Sara C Shalin, and Alan J Tackett. “Current state of melanoma diagnosis and treatment.” In: Cancer biology & therapy 20.11 (2019), pp. 1366–1379.

[19] Fabiola De Marco, Filomena Ferrucci, Michele Risi, and Genoveffa Tortora. “Classification of QRS complexes to detect Premature Ventricular Contraction using machine learning techniques.” In: Plos one 17.8 (2022), e0268555.

[20] Fabiola De Marco, Dewar Finlay, and Raymond R Bond. “Classification of Premature Ventricular Contraction Using Deep Learning.” In: *2020* Computing in Cardiology. IEEE. 2020, pp. 1–4.

[21] Luigi Di Biasi, Alessia Auriemma Citarella, Michele Risi, and Genoveffa Tortora. “A Cloud Approach for Melanoma Detection Based on Deep Learning Networks.” In: IEEE Journal of Biomedical and Health Informatics 26.3 (2021), pp. 962–972.

[22] Luigi Di Biasi, Fabiola De Marco, Alessia Auriemma Citarella, Paola Barra, Stefano Piotto Piotto, and Genoveffa Tortora. “Hybrid approach for the design of CNNs using Genetic Algorithms for Melanoma Classification.” In: AHIA *2022* (2022).

[23] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey.” In: *2018 41*st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE. 2018, pp. 0210–0215.

[24] Gianni D’Angelo and Francesco Palmieri. “GGA: A modified genetic algorithm with gradient-based local search for solving constrained optimization problems.” In: Information Sciences 547 (2021), pp. 136–162.

[25] Luigi D’Arco, Haiying Wang, and Huiru Zheng. “Assessing Impact of Sensors and Feature Selection in Smart-Insole-Based Human Activity Recognition.” In: Methods and Protocols 5.3 (2022), p. 45.

[26] Mila Efimenko, Alexander Ignatev, and Konstantin Koshechkin. “Review of medical image recognition technologies to detect melanomas using neural networks.” In: BMC bioinformatics 21.11 (2020), pp. 1–7.

[27] Adewale Fadaka, Basiru Ajiboye, Oluwafemi Ojo, Olusola Adewale, Israel Olayide, and Rosemary Emuowhochere. “Biology of glucose metabolization in cancer cells.” In: Journal of Oncological Sciences 3.2 (2017), pp. 45–51.

[28] Y Fujisawa, Y Otomo, Y Ogata, Y Nakamura, R Fujita, Y Ishitsuka, R Watanabe, N Okiyama, K Ohara, and M Fujimoto. “Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis.” In: British Journal of Dermatology 180.2 (2019), pp. 373–381.

[29] P Gaviani, ME Mullins, TA Braga, ET Hedley-Whyte, Elkan F Halpern, PS Schaefer, and John W Henson. “Improved detection of metastatic melanoma by T2*-weighted imaging.” In: American journal of neuroradiology 27.3 (2006), pp. 605–608.

[30] Jeffrey E Gershenwald, Richard A Scolyer, Kenneth R Hess, Vernon K Sondak, Georgina V Long, Merrick I Ross, Alexander J Lazar, Mark B Faries, John M Kirkwood, Grant A McArthur, et al. “Melanoma staging: evidence-

based changes in the American Joint Committee on Cancer eighth edition cancer staging manual." In: *CA: a cancer journal for clinicians* 67.6 (2017), pp. 472–492.

[31] Barbara A Gilchrest, Mark S Eller, Alan C Geller, and Mina Yaar. "The pathogenesis of melanoma induced by ultraviolet radiation." In: *New England Journal of Medicine* 340.17 (1999), pp. 1341–1348.

[32] Stephen Gilmore, Rainer Hofmann-Wellenhof, and H Peter Soyer. "A support vector machine for decision support in melanoma recognition." In: *Experimental Dermatology* 19.9 (2010), pp. 830–835.

[33] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images." In: *Expert systems with applications* 42.19 (2015), pp. 6578–6585.

[34] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities." In: *Computers in Biology and Medicine* 127 (2020), p. 104065.

[35] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. "XAI—Explainable artificial intelligence." In: *Science robotics* 4.37 (2019), eaay7120.

[36] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)." In: *arXiv preprint arXiv:1605.01397* (2016).

[37] Sarah Haggenmüller, Roman C Maron, Achim Hekler, Jochen S Utikal, Catarina Barata, Raymond L Barnhill, Helmut Beltraminelli, Carola Berking, Brigid Betz-Stablein, Andreas Blum, et al. "Skin cancer classification via convolutional neural networks: systematic review of studies

involving human experts." In: European Journal of Cancer 156 (2021), pp. 202–216.

[38] MF Healsmith, JF Bourke, JE Osborne, and RAC Graham-Brown. "An evaluation of the revised seven-point checklist for the early diagnosis of cutaneous malignant melanoma." In: British Journal of Dermatology 130.1 (1994), pp. 48–50.

[39] Walter D Holder Jr, Richard L White Jr, James H Zuger, Edward J Easton Jr, and Frederick L Greene. "Effectiveness of positron emission tomography for the detection of melanoma metastases." In: Annals of surgery 227.5 (1998), p. 764.

[40] Khalid M Hosny, Mohamed A Kassem, and Mohamed M Foaud. "Classification of skin lesions using transfer learning and augmentation with Alex-net." In: PloS one 14.5 (2019), e0217293.

[41] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. cite arxiv:1704.04861. 2017. url: http://arxiv.org/abs/1704.04861.

[42] Xing Hu, Ge Li, Xin Xia, David Lo, and Zhi Jin. "Deep code comment generation." In: *2018* IEEE/ACM *26*th International Conference on Program Comprehension (ICPC). IEEE. 2018, pp. 200–20010.

[43] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks." In: *2017* IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[44] Hsin-Wei Huang, Benny Wei-Yun Hsu, Chih-Hung Lee, and Vincent S Tseng. "Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers." In: The Journal of Dermatology 48.3 (2021), pp. 310–316.

[45] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with *50*x fewer parameters and *<0.5*MB model size. cite arxiv:1602.07360Comment: In ICLR Format. 2016. url: http://arxiv.org/abs/1602.07360.

[46] Shunichi Jinnai, Naoya Yamazaki, Yuichiro Hirano, Yohei Sugawara, Yuichiro Ohe, and Ryuji Hamamoto. “The development of a skin cancer classification system for pigmented skin lesions using deep learning.” In: Biomolecules 10.8 (2020), p. 1123.

[47] Mohammad Ali Kadampur and Sulaiman Al Riyaee. “Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images.” In: Informatics in Medicine Unlocked 18 (2020), p. 100282. issn: 2352-9148. doi: https://doi.org/10.1016/j.imu.2019.100282. url: https://www.sciencedirect.com/science/article/pii/S2352914819302047.

[48] Hisashi Kanemaru, Yukari Mizukami, Akira Kaneko, Ikko Kajihara, and Satoshi Fukushima. “Promising blood-based biomarkers for Melanoma: Recent progress of liquid Biopsy and its future perspectives.” In: Current Treatment Options in Oncology (2022), pp. 1–16.

[49] Sourabh Katoch, Sumit Singh Chauhan, and Vijay Kumar. “A review on genetic algorithm: past, present, and future.” In: Multimedia Tools and Applications 80.5 (2021), pp. 8091–8126.

[50] Ranpreet Kaur, Hamid GholamHosseini, Roopak Sinha, and Maria Lindén. “Melanoma Classification Using a Novel Deep Convolutional Neural Network with Dermoscopic Images.” In: Sensors 22.3 (2022), p. 1134.

[51] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. “Seven-point checklist and skin lesion classification using multitask multimodal neural nets.” In: IEEE journal of biomedical and health informatics 23.2 (2018), pp. 538–546.

[52] Minoru Kitago, Kazuo Koyanagi, Takeshi Nakamura, Yasufumi Goto, Mark Faries, Steven J O'day, Donald L Morton, Soldano Ferrone, and Dave SB Hoon. "mRNA expression and BRAF mutation in circulating melanoma cells isolated from peripheral blood with high molecular weight melanoma-associated antigen-specific monoclonal antibody beads." In: Clinical chemistry 55.4 (2009), pp. 757–764.

[53] Oliver Kramer. "Genetic algorithms." In: Genetic algorithm essentials. Springer, 2017, pp. 11–19.

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: Advances in neural information processing systems. 2012, pp. 1097–1105. url: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[55] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: Communications of the ACM 60.6 (2017), pp. 84–90.

[56] Manoj Kumar, Dr Husain, Naveen Upreti, Deepti Gupta, et al. "Genetic algorithm: Review and application." In: Available at SSRN *3529843* (2010).

[57] Annu Lambora, Kunal Gupta, and Kriti Chopra. "Genetic algorithm-A literature review." In: *2019* international conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE. 2019, pp. 380–384.

[58] Tim Lee, Vincent Ng, Richard Gallagher, Andrew Coldman, and David McLean. "Dullrazor®: A software approach to hair removal from images." In: Computers in biology and medicine 27.6 (1997), pp. 533–543.

[59] Howard F Marx, Patrick M Colletti, Janak K Raval, William D Boswell Jr, and Chi-Shing Zee. "Magnetic resonance imaging features in melanoma." In: Magnetic resonance imaging 8.3 (1990), pp. 223–229.

[60] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics." In: Briefings in bioinformatics 18.5 (2017), pp. 851–869.

[61] Saumitra Mishra, Bob L Sturm, and Simon Dixon. "Local interpretable model-agnostic explanations for music content analysis." In: ISMIR. Vol. 53. 2017, pp. 537–543.

[62] Kajsa Møllersen, Herbert Kirchesch, Maciel Zortea, Thomas R Schopf, Kristian Hindberg, and Fred Godtliebsen. "Computer-aided decision support for melanoma detection applied on melanocytic and nonmelanocytic skin lesions: a comparison of two systems based on automatic analysis of dermoscopic images." In: BioMed Research International 2015 (2015).

[63] Gerd Muehllehner and Joel S Karp. "Positron emission tomography." In: Physics in Medicine & Biology 51.13 (2006), R117.

[64] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. "The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions." In: Journal of the American Academy of Dermatology 30.4 (1994), pp. 551–559.

[65] Ebrahim Nasr-Esfahani, Shadrokh Samavi, Nader Karimi, S Mohamad R Soroushmehr, Mohammad H Jafari, Kevin Ward, and Kayvan Najarian. "Melanoma detection by analysis of clinical images using convolutional neural network." In: *2016 38*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE. 2016, pp. 1373–1376.

[66] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms." In: IEEE Transactions on Systems, Man, and Cybernetics 9.1 (1979), pp. 62–66. doi: 10.1109/TSMC.1979.4310076.

[67] Nobuyuki Otsu. "A threshold selection method from gray-level histograms." In: IEEE transactions on systems, man, and cybernetics 9.1 (1979), pp. 62–66.

[68] Sumit Paliwal, Byeong Hee Hwang, Kenneth Y Tsai, and Samir Mitragotri. “Diagnostic opportunities based on skin biomarkers.” In: European journal of pharmaceutical sciences 50.5 (2013), pp. 546–556.

[69] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning.” In: IEEE Transactions on knowledge and data engineering 22.10 (2009), pp. 1345–1359.

[70] Eduardo Pérez and Sebastián Ventura. “An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis.” In: Neural Computing and Applications (2021), pp. 1–20.

[71] S Perumal and T Velmurugan. “Preprocessing by contrast enhancement techniques for medical images.” In: International Journal of Pure and Applied Mathematics 118.18 (2018), pp. 3681–3688.

[72] Tri-Cong Pham, Chi-Mai Luong, Van-Dung Hoang, and Antoine Doucet. “AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function.” In: Scientific Reports 11.1 (2021), pp. 1–13.

[73] Stefano Piotto, Luigi Di Biasi, Simona Concilio, Aniello Castiglione, and Giuseppe Cattaneo. “GRIMD: distributed computing for chemists and biologists.” In: Bioinformation 10.1 (2014), p. 43.

[74] Dan Popescu, Mohamed El-Khatib, Hassan El-Khatib, and Loretta Ichim. “New Trends in Melanoma Detection Using Neural Networks: A Systematic Review.” In: Sensors 22.2 (2022), p. 496.

[75] Li-na Qi, Bo Zhang, and Zhan-kai Wang. “Application of the Otsu method in image processing.” In: Radio Engineering of China 7.009 (2006).

[76] Maryam Ramezani, Alireza Karimian, and Payman Moallem. “Automatic detection of malignant melanoma using macroscopic images.” In: Journal of medical signals and sensors 4.4 (2014), p. 281.

[77] Marco Rastrelli, Saveria Tropea, Carlo Riccardo Rossi, and Mauro Alaibac. “Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification.” In: In vivo 28.6 (2014), pp. 1005–1011.

[78] Navid Razmjooy, B Somayeh Mousavi, Fazlollah Soleymani, and M Hosseini Khotbesara. “A computer-aided diagnosis system for malignant melanomas.” In: Neural Computing and Applications 23.7 (2013), pp. 2059–2071.

[79] Aaron N. Richter and Taghi M. Khoshgoftaar. “Efficient learning from big data for cancer risk modeling: A case study with melanoma.” In: Computers in Biology and Medicine 110 (2019), pp. 29–39. issn: 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2019.04.039. url: https://www.sciencedirect.com/science/article/pii/S0010482519301477.

[80] Daniel Ruiz, Vicente Berenguer, Antonio Soriano, and Belén Sánchez. “A decision support system for the diagnosis of melanoma: A comparative approach.” In: Expert Systems with Applications 38.12 (2011), pp. 15217–15223.

[81] Martina Sanlorenzo, Igor Vujic, Christian Posch, Akshay Dajee, Adam Yen, Sarasa Kim, Michelle Ashworth, Michael D Rosenblum, Alain Algazi, Simona Osella-Abate, et al. “Melanoma immunotherapy.” In: Cancer biology & therapy 15.6 (2014), pp. 665–674.

[82] A Hunter Shain and Boris C Bastian. “From melanocytes to melanomas.” In: nature reviews Cancer 16.6 (2016), pp. 345–358.

[83] K Aditya Shastry and H A Sanjay. “Machine learning for bioinformatics.” In: Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications. Springer, 2020, pp. 25–39.

[84] Takahiro Shinozaki and Shinji Watanabe. “Structure discovery of deep neural network based on evolutionary algorithms.” In: *2015* IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2015, pp. 4979–4983.

[85] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning.” In: Journal of big data 6.1 (2019), pp. 1–48.

[86] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. cite arxiv:1409.1556. 2014. url: http://arxiv.org/abs/1409.1556.

[87] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking the Inception Architecture for Computer Vision.” In: *2016* IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308. url: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.308.

[88] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. cite arxiv:1409.4842. 2014. url: http://arxiv.org/abs/1409.4842.

[89] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going deeper with convolutions.” In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 1–9.

[90] Arthur Tenenhaus, Alex Nkengne, Jean-François Horn, Camille Serruys, Alain Giron, and Bernard Fertil. “Detection of melanoma from dermoscopic images of naevi acquired under uncontrolled conditions.” In: Skin Research and Technology 16.1 (2010), pp. 85–97.

[91] J Meirion Thomas. “Prognostic false-positivity of the sentinel node in melanoma.” In: Nature Clinical Practice Oncology 5.1 (2008), pp. 18–23.

[92] S Vishnu, SR Jino Ramson, and R Jegan. “Internet of medical things (IoMT)-An overview.” In: *2020* 5th international conference on devices, circuits and systems (ICDCS). IEEE. 2020, pp. 101–104.

[93] Charles M Washington and Dennis T Leaver. Principles and Practice of Radiation Therapy-E-Book. Elsevier Health Sciences, 2015.

[94] Caroline G Watts, Anne E Cust, Scott W Menzies, Elliot Coates, Graham J Mann, and Rachael L Morton. “Specialized surveillance for individuals at high risk for melanoma: a cost analysis of a high-risk clinic.” In: JAMA dermatology 151.2 (2015), pp. 178–186.

[95] Martin A Weinstock. “Early detection of melanoma.” In: Jama 284.7 (2000), pp. 886–889.

[96] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning.” In: Journal of Big data 3.1 (2016), pp. 1–40.

[97] WW Woodruff Jr, WT Djang, RE McLendon, ER Heinz, and DR Voorhees. “Intracerebral malignant melanoma: high-field-strength MR imaging.” In: Radiology 165.1 (1987), pp. 209–213.

[98] Xueli Xiao, Ming Yan, Sunitha Basodi, Chunyan Ji, and Yi Pan. “Efficient hyperparameter optimization in deep learning using a variable length genetic algorithm.” In: arXiv preprint arXiv:*2006.12703* (2020).

[99] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. “Automated melanoma recognition in dermoscopy images via very deep residual networks.” In: IEEE transactions on medical imaging 36.4 (2016), pp. 994–1004.

[100] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. “Attention residual learning for skin lesion classification.” In: IEEE transactions on medical imaging 38.9 (2019), pp. 2092–2103.

[101] Tianhu Zhang, Yuanjun Liu, Yandi Rao, Xiaopeng Li, and Qingxin Zhao. “Optimal design of building environment with hybrid genetic algorithm, artificial neural network, multivariate regression analysis and fuzzy logic controller.” In: Building and Environment 175 (2020), p. 106810.

[102] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6848–6856. doi: 10.1109/CVPR.2018.00716.

[103] Xiaoqing Zhang. "Melanoma segmentation based on deep learning." In: *Computer Assisted Surgery* 22.sup1 (2017), pp. 267–277.

[104] Yu Zhou and Zhuoyi Song. "Binary decision trees for melanoma diagnosis." In: *Proceedings of the International Workshop on Multiple Classifier Systems*. Springer. 2013, pp. 374–385.