Universitat de Girona
Escola Politècnica Superior

Final Degree Project

Grade: Biomedical Engineering

Title: Early detection of Melanoma using Artificial Intelligence

Document: Memory

Student: Iván González González

Tutor: Rafael Garcia Campos

Department: Arquitectura i Tecnologia de Computadors

Area: Arquitectura i Tecnologia de Computadors

Call (month/year): September 2024

Acknowledgments

I would like to use this section to thank all the people who have supported me during this exciting stage of my life.

Firstly, I would like to thank my family, especially my parents, my brother, and my grandmothers, for giving me all their love and support in every step I have taken. I am convinced that, without them, none of this would have been possible. I would also like to remember the people who saw me start this path and who unfortunately will not see me finish it, but I am sure they would be proud of me: my grandparents, Mavi, I love you, I have finally achieved it, I am an engineer.

Of course, I also want to thank all my friends and colleagues who, during these long 5 years, have been by my side, shoulder to shoulder, supporting me. The experience has been much more beautiful and cooler with all of you.

I would also like to thank all the people in the ViCOROB group for their patience and for always being there to help me or resolve any questions I had. Especially, of course, to my tutor, Rafa. Without him, this work would not have been possible. Thank you for giving me the opportunity to work with you on this project and for guiding and helping me in it. I have learned much more than I could have imagined, and I have loved working with you. I am convinced that you are the teacher that every student would dream to have. Many thanks for everything.

This achievement would not have been possible without the support of all these people. Both those that have been with me from the first moment and those that have appeared along the way. Despite the difficulties, I have achieved it.

Thank you very much, from the bottom of my heart, to everyone.

Acronyms

Although the number of acronyms in this project is small, it is important to mention them to have a better lecture of the document. These are the acronyms used:

- **AI** = Artificial Intelligence
- **CNN** = Convolutional neural network
- **ML** = Machine learning
- **DL** = Deep learning
- **Crops** = Lesion images cropped from full body maps images extracted by 3D TBP scanners.
- **FLOPS** = Floating point operations per second (FLOPS)
- **OS** = Oversampling
- **US** = Undersampling
- **TP** = True positive
- **TN** = True negative
- **FP** = False positive
- **FN** = False negative

Index

Chapter 1

Introduction

Skin cancer, particularly melanoma, is one of the most aggressive forms of cancer and has lately been raising the incidence rate globally. Despite the advancements in medical imaging and dermatological practices, early detection is still a big challenge that often leads to late diagnosis and poor patient outcomes. Melanoma, if identified in its early stages, is highly treatable; however, once it progresses, the survival rate drops drastically. This shows us how important and critically improved diagnostic methods that can help clinicians detect melanoma at its earliest stages are needed.

During the last few years, AI has gained a big role as a transformative technology in various fields, including, in this case, healthcare. AI's ability to analyze enormous amounts of data and identify patterns that may not be as apparent to the human eye has made it a very valuable tool in medical diagnostics. A tool that, from my point of view, won't make clinicians disappear but will create a powerful gap between those who use it and those who do not in their day-to-day work. Machine learning has shown a promising capacity for the task of analyzing medical images and may have the opportunity to bring clinicians potential solutions to these challenges in early cancer detection.

The aim of this project is to explore the applications of AI, especially deep learning and machine learning, in the detection of melanoma in crops. This objective will try to be accomplished by using advanced CNNs and transfer learning techniques to develop a model that can be used to assist dermatologists in identifying malignant skin lesions with a high accuracy. Patient metadata, such as age, gender, anatomical site and others, will be used in the training of some models to see if this data helps the model in its predictive capabilities. Genomical data has finally not been used due to some technical issues and availability issues.

This project is built to not only create an effective diagnostic tool, but also to contribute to the growing knowledge of the use of AI in this field. With rigorous testing and validation, this research will try to demonstrate the potential of AI to help improve clinical outcomes in melanoma and skin cancer detection. The project will be developed in Python using PyTorch as the selected ML and DL framework.

Chapter 2

Previous concepts

In this chapter all the necessary concepts will be introduced for a better understanding and a good follow-up of this project.

2.1. Skin cancer and melanoma

Malignant melanoma is a serious type of skin cancer that develops when melanocytes (the cells that give the skin its tan or brown color) start to grow out of control. Skin cancer mostly occurs in three main forms: basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and melanoma, although there are other rare types that can occur.

Melanoma is much less common than these other types of skin cancers, but is more dangerous because it's much more likely to spread to other parts of the body (causing metastasis) if not found and treated early. Here is where AI models/tools such as the one developed in this project can be helpful for the clinicians for early detecting melanoma.

Melanocytes are skin cells found in the upper layer of the skin. They produce a pigment known as melanin, which gives skin its color. There are two types of melanin, eumelanin and pheomelanin. When our skin is exposed to ultraviolet (UV) radiation from the sun or tanning beds, it causes skin damage that triggers melanocytes to produce more melanin, but only the eumelanin pigment attempts to protect the skin by causing the skin to darken or tan. Melanoma occurs when DNA damage from sunburns or tanning due to UV radiation triggers changes (mutations) in the melanocytes, resulting in, as said before, uncontrolled cellular growth.

Naturally darker-skinned people have more eumelanin while fair-skinned people have more pheomelanin. While eumelanin has the ability to protect the skin from sun damage, pheomelanin does not. This is the reason why people with darker skin have a lower risk of developing skin cancer. Fair-skinned people, due to the lack of eumelanin, are more susceptible to sun damage, burning and skin cancer. Even so, skin cancer can happen to anyone, regardless of their skin tone.

Skin cancer types

Most skin cancers start in the top layer of the skin, called the epidermis. In this layer, we can find three main types of cells:

- **Squamous cells:** These are flat cells in the upper (outer) part of the epidermis, which are constantly shed as new ones form.
- **Basal cells:** These cells are in the lower part of the epidermis, called the basal cell layer. These cells constantly divide to form new cells to replace the squamous cells that wear off the skin's surface. As these cells move up in the epidermis, they get flatter, eventually becoming squamous cells.
- **Melanocytes:** These are the cells that can become melanoma. They normally make a brown pigment called melanin, which gives the skin its tan or brown color. Melanin protects the deeper layers of the skin from some of the harmful effects of the sun.

The epidermis is separated from the deeper layers of skin by the basement membrane. When a skin cancer becomes more advanced, it generally grows through this barrier and into the deeper layers.

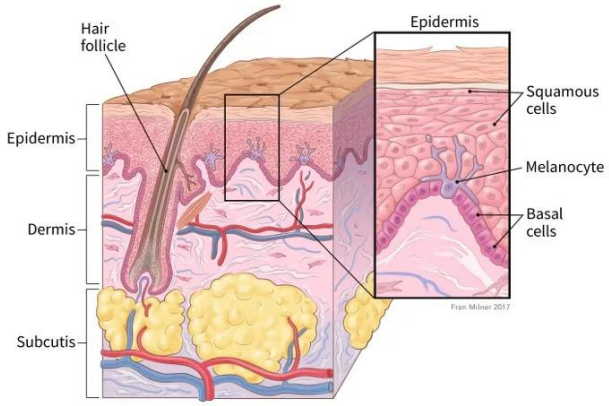

Figure 1: Skin anatomical structure schema.

Symptoms

Some melanomas develop from existing moles. The rest grow on what was previously normal skin. It's important to know what your skin looks like normally. This way, you will notice any unusual changes in any moles or skin zones.

Clinicians usually use a checklist, which explains some of the signs of melanoma that people should look out for. It's called the ABCDE list as can be seen in Figure 2. Each of the letters refers to a characteristic.

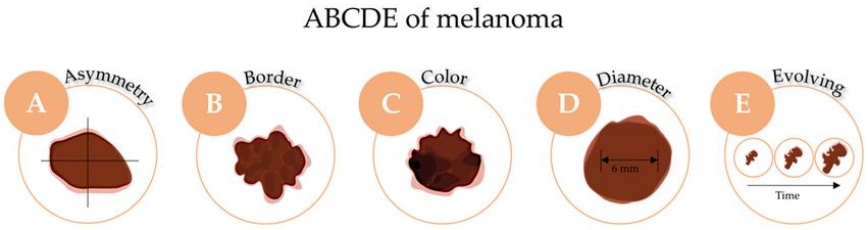

Figure 2: ABCDE of melanoma.

The meaning of each letter is:

- **A - asymmetrical:** Shape of the mole or abnormal patch of skin. Melanomas are likely to have an uneven shape. Both halves may have different shapes or sizes (asymmetrical). Normal moles tend to have a more even shape, and both halves are usually similar (more or less symmetrical).
- **B - border:** Edges of the mole or abnormal patch of skin. Melanomas are likely to have irregular edges (borders). Normal moles usually have a smooth and regular border.
- **C - color:** Melanomas are often an uneven color and contain more than one color shade. Normal moles usually have an even color. And even if they have two colors in them, these are normally symmetrical across the two halves of the melanoma.
- **D - diameter:** Most melanomas are more than 6mm wide. But they can be smaller if they are diagnosed early. Normal moles are usually about the size of the end of a pencil or smaller. From 3+ mm onwards, it might be a good idea to have the mole revised by a doctor just in case.
- **E - evolving:** Evolving means changing. Melanomas might change in size, shape or color. Other changes to be noticed can be bleeding, itching, a change in sensation to the mole or area, or the mole becoming crusty.

Melanoma in people with brown or black skin may be more difficult to see and may show any of the ABCD signs.

Risks and causes

The risk of developing melanoma depends on many things, including lifestyle factors and some medical conditions.

- **Age:** The risk of melanoma increases with age, so it's more common in older people.
- **Ultraviolet (UV) light:** UV light is the main environmental factor that increases the risk of developing melanoma. UV light comes from the sun or sunbeds. So in this section, we can include sun exposure, sunburn and sunbeds.
- **Skin color and freckling:** People who have white skin, especially those with fair or red hair, are more at risk of developing melanoma. So are people with lots of freckles. People with black or brown skin can still get melanoma, but they have more natural protection against it, as explained before.

- **Mole quantity:** The more moles you have on your body, the higher your risk of melanoma is. There are some studies that found that melanoma risk was higher in people with one hundred or more common moles. People who have lots of unusually shaped or large moles (>5mm across) have a higher risk of melanoma than the general population.
- **Birthmarks:** Birthmarks are colored marks on a baby's skin when it's born. They can develop as well in the few weeks after birth. Different types of cells make up different types of birthmarks. Most birthmarks, such as the common port wine stains and strawberry marks, carry no risk of developing into cancer. But a very rare type, called a giant congenital melanocytic naevus, can develop into melanoma if it's larger than twenty centimeters. Doctors recommend that you check all birthmarks regularly for any signs of change. People who have a large congenital melanocytic naevus should have regular checks by a skin specialist (dermatologist).
- **Family history and genetic factors:** Your risk of melanoma is higher if you have a close relative who's also had melanoma in the past. This is probably partly because we tend to share the same coloring and skin type as our close relatives. Your risk is highest if your relative had melanoma when they were younger than thirty or more than one first degree relative had melanoma. Genetic factors play an important part here, around 10% of the cases of melanoma might be linked to an inherited gene change. There are a number of genes that might be associated with increased risk. This gene is called CDKN2A, known to cause FAMMM (familial atypical multiple mole melanoma syndrome). People with FAMMM have more than fifty moles, some of which are atypical and often different-sized. And at least one close relative who has had melanoma. For the small number of families who carry these genes, sun protection is even more important.
- **Other medical conditions**, such as different diseases or weakened immune systems, can make you more likely to develop melanoma.
- **Chemicals in the workspace:** Polychlorinated biphenyls (PCBs) are known to increase the risk of melanoma skin cancer. These chemicals are often found in some old electrical equipment.

Types of melanomas

Melanomas can start anywhere on the skin, but in people with lighter skin color, they are more likely to start on the trunk (chest and back) in men and on the legs in women. The neck and face are also other common sites.

There are different types of skin melanoma. The most common types are:

- **Superficial spreading melanoma:** Most common type of melanoma. Usually flat and irregular in form and color, with black/brown coffee variable shades. Grows across the skin before growing deeper down into it. It's more common on lighter skin color people.
- **Nodular melanoma:** Tend to grow downward into deeper layers of skin. They can grow quite quickly. It usually starts with a blue-black or red-blue raised area on the skin surface. Some cannot have any color. It is the second-most common type of melanoma.

- **Lentigo maligna melanoma:** The affected area is flat and gets might slowly get bigger over several years and can change shape or color. It usually appears in elderly people. They appear in areas of skin that get a lot of sun exposure, so predominantly on the head/neck. More common in people who have spent a lot of time in the sun.
- **Acral lentiginous melanoma:** Rarely, melanoma can also develop in areas not exposed to the sun. Melanomas in the palms of the hands, soles of the feet, and under the nails are normally a type of melanoma called acral melanomas. They are not related to exposure to the sun. Acral melanomas are more commonly diagnosed in people with brown or black skin and middle-aged and older adults.
- **Ocular melanomas:** Melanomas can also form in other parts of the body, such as the inside of the eye. The most common type of melanoma of the eye starts in the uvea, the middle layer of the eye.
- **Mucosal melanomas:** Melanoma that starts in the mucous membrane is called mucosal melanoma. Although it's quite rare, it can develop inside the nose, mouth, throat, genital or anal area, and digestive system. These cases are much less common than melanoma of the skin.
- **Other rare melanoma skin cancer types:** These include spitzoid melanoma, malignant blue nevus and amelanotic melanoma.

There are many other types of skin cancer. Skin cancers that are not melanomas are sometimes grouped as non-melanoma skin cancers because they develop from other skin cells different from melanocytes. They tend to behave very differently from melanomas and are often treated with different methods. Basal cell cancer (BCC) and squamous cell cancer (SCC) are by far the most common types of skin cancer. These cancers (especially BCCs) are much less likely to spread (metastasize) to other parts of the body than are melanomas, so they are usually less concerning and are treated differently.

Stages

The TNM staging system is how the doctor stages your melanoma skin cancer. The stage of a cancer tells the doctor how thick it is and how far it's spreading, so a treatment for your need can be chosen. TNM stands for Tumor, Node, Metastasis.

<u>Tumor (T)</u>
Tumor describes the thickness of the melanoma. There are six main stages of tumor thickness in melanoma. These go from Tis to T4.

- Tis: Melanoma cells are only in the very top layer of the skin surface.
- T0: Means no melanoma cells can be seen where the melanoma started.
- T1: Means the melanoma is 1mm thick or less. It is actually divided into two subtypes, T1a and T1b.
    - T1a means the melanoma is less than 0.8 mm thick and the skin over the tumor does not look broken under a microscope (not ulcerated).

- T1b means the melanoma is less than 0.8 mm thick but is ulcerated, or that the melanoma is between 0.8 mm and 1.0 mm and may or may not be ulcerated.

- T2: Means the melanoma is between 1 mm and 2 mm thick.
- T3: Means the melanoma is between 2 mm and 4 mm thick.
- T4: Means the melanoma is more than 4 mm thick.

T2 and T4 are also divided into a and b depending on whether it is ulcerated or not, same procedure as T1 stage.

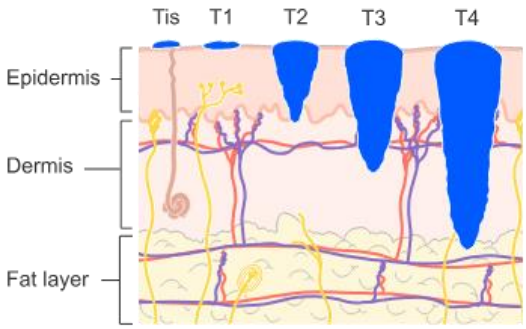

Figure 3: Tumor (T) staging of melanoma.

Node (N)

The node stage describes whether the cancer cells are in the nearby lymph nodes. There are four main stages in melanoma.

Doctors also look at whether there are areas of cancer spread (metastases) between the primary tumor and the nearby lymph nodes. They use different terms for this:

- Microsatellite metastases
- Satellite metastases
- In-transit metastases

Microsatellite metastases mean that, under a microscope, a melanoma cell can be seen next to the primary tumor.

Satellite metastases mean that the melanoma cells have visibly spread to an area less than 2 cm away from the primary tumor.

In-transit metastases mean the melanoma has spread to an area more than 2 cm away from the primary tumor but before the nearby lymph nodes.

- N0 means there are no melanoma cells in the nearby lymph nodes.
- N1 means there are melanoma cells in one lymph node or there are in-transit, satellite or microsatellite metastases.

- N2 means there are melanoma cells in 2 or 3 lymph nodes or there are melanoma cells in one lymph node, and there are also in-transit, satellite or microsatellite metastases.
- N3 means there are melanoma cells in 4 or more lymph nodes or there are melanoma cells in 2 or 3 lymph nodes and there are in-transit, satellite or microsatellite metastases or there are melanoma cells in any number of lymph nodes and they have stuck to each other (matted lymph nodes).

Metastasis (M)

Metastasis (M) describes whether the cancer has spread to a different part of the body.

There are two stages of metastasis: M0 and M1.

- M0 means the cancer hasn't spread to another part of the body.
- M1 means the cancer has spread to another part of the body.

M1 can be further divided depending on which parts of the body the cancer has spread to and whether there are raised levels of a chemical in the blood called lactate dehydrogenase (LDH).

Benign skin tumors

Many types of benign (non-cancerous) tumors can develop from different types of skin cells.

A mole (a nevus) is a benign skin tumor that develops from melanocytes. Almost everyone has some moles. Nearly all moles (nevi) are harmless, but having some types can raise your risk of melanoma.

A Spitz nevus is a kind of mole that sometimes looks like melanoma. It's more common in children and teens, but it can also be seen in adults. These tumors are typically benign and don't spread. But sometimes doctors have trouble telling Spitz nevi from true melanomas, even when looking at them under a microscope. Therefore, they are often removed, just to be safe.

Benign tumors that develop from other types of skin cells

- Seborrheic keratoses
- Hemangiomas
- Lipomas
- Warts

Treatment

The stage of your cancer helps to decide what treatment suits your needs. The treatment will also depend on:

- The zone where the melanoma is located.
- Your health and level of fitness.

Surgery is the main treatment for people with melanoma that has not spread (early melanoma). You usually have other treatment if your melanoma is at high risk of coming back, unable to have surgery, or has spread to another part of the body (advanced). Other treatments include:

- targeted cancer drug
- immunotherapy
- chemotherapy
- radiotherapy
- laser therapy
- cryotherapy
- taking part in a clinical trial

2.2 Introduction to AI

Artificial intelligence is a field of science whose objective is to build computers and machines that can reason, learn, and act in a way that would usually require human intelligence or that involves an enormous amount of data whose scale exceeds what we humans can analyze.

AI is a very big field that has many different disciplines in it, including computer science, data analytics and statistics, hardware and software engineering, linguistics, neuroscience, and even philosophy and psychology.

How does AI work?

Although the specifics vary across different AI techniques, the core principle revolves around the data used to train it. AI systems can learn and improve through being exposed to very big amounts of data, to identify patterns and relationships that humans may miss.

During these learning processes, algorithms are often involved, which are no other than sets of rules or instructions that guide the AI's analysis and decision-making. In ML, these algorithms are trained on labeled or unlabeled data to make predictions or categorize information.

DL, a more specialized field, utilizes artificial neural networks with multiple layers to process information, with the aim of mimicking the structure and function of the human brain in decision-making. Through continuous learning and adaptation, AI systems become increasingly adept at performing specific tasks, from recognizing images to translating languages and beyond.

Types of AI

AI can be organized in many different ways, depending on stages of development or actions performed.

These are the four stages of AI development commonly recognized:

1. **Reactive machines:** Limited AI that only reacts to different kinds of stimulus based on a set of preprogrammed rules. Does not use memory and can not learn with new data.
2. **Limited memory:** Most modern AI systems are considered to have limited memory. It can use memory to improve over time by being trained with new data, typically through an artificial neural network or any other training model. DL for example is considered to be limited memory AI.
3. **Theory of mind:** Theory of mind AI does not currently exist, but there is research ongoing into its possibilities. It would be the case of an AI that can emulate the human mind and has decision-making capabilities equal to a human. This includes recognition and remembering emotions and reaction in social situations as a human would do.
4. **Self-aware:** In this case, we are talking about a step above the theory of mind AI. Self-aware AI describes the mythical machine that is aware of its own existence and has the intellectual and emotional capabilities of a human. Self-aware AI does not currently exist.

A lot of times, the types of artificial intelligence are categorized by what the machine can do. At this moment, all of what we currently call artificial intelligence is considered commonly artificial "narrow" intelligence, it can only perform a certain set of actions based on its programming and training. Artificial general intelligence (AGI) would be the ability for a machine to "sense, think, and act" just like a human. It does not currently exist. The next level would be artificial superintelligence (ASI), in which the machine would be able to function in all ways superior to a human.

2.2.1. Machine Learning

ML is a field of study in AI that consists on the development and study of statistical algorithms that can make predictions or decisions based on data. It covers a large number of techniques that enable computers to learn from and make decisions based on data without being explicitly programmed for specific tasks.

There are many types of machine learning techniques or algorithms. Some of the most known are linear regression, logistic regression, decision trees, random forest, support vector machines (SVMs), k-nearest neighbor (KNN), clustering and more. Each one of these approaches mentioned is suited to different kinds of problems and data.

ML finds application in many fields, including natural language processing, computer vision, speech recognition, business, email filtering, agriculture and medicine. One of the most popular types of ML algorithms are called neural networks (or artificial neural networks).

2.2.2. Deep learning and neural networks

DL is a subfield of machine learning that uses artificial neural networks with multiple layers, also known as deep neural networks, to learn hierarchical representations of data. Neural networks are inspired by the structure of the human brain, specifically the interconnected network of neurons that we have. Due to this structure of using the connections between the artificial neurons, these algorithms are capable of learning, predicting and managing enormous quantities of data.

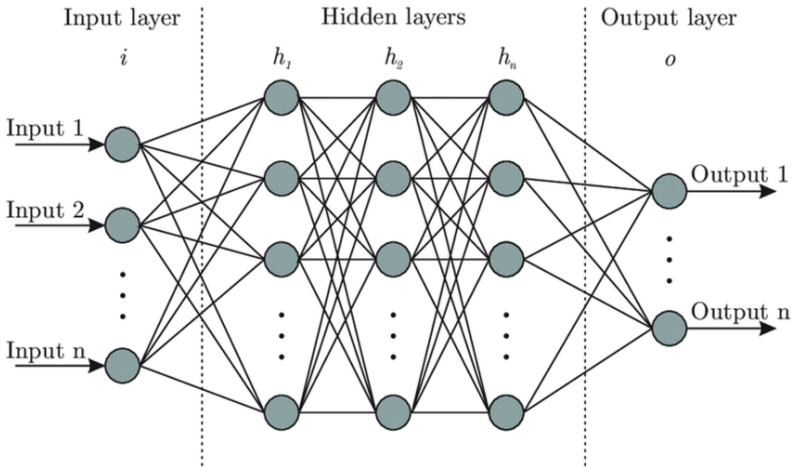

Figure 4. Structure of deep learning scheme

What are artificial neurons?

The basic computational unit of the brain is a neuron. In the case of DL algorithms, the basic units are artificial neurons also known as perceptrons. Their function is to try to mimic and simulate how a real neuron would work. By creating connections between these perceptrons, we get an artificial neural network, a structure we use to create models capable of taking decisions and making predictions by their own.

How do artificial neurons work?

Initially, the neurons receive entries. Once all entries are received, each entry gets multiplied by their correspondent weight. Next, a sum of all products previously done is realized and the results go through an activation function. This activation function finally gives us what the output of the neuron will be, a scheme can be seen in Figure 5.

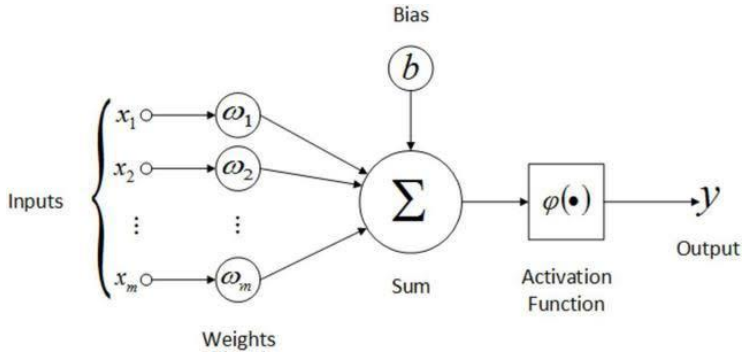

Figure 5. Functioning scheme of an artificial neuron (perceptron).

What is a activation function?

An activation function is a function that, from an entry and a mathematical function, returns us an output. These outputs are usually between a determined range of values, usually from 0 to 1 or from -1 to 1.

Activation functions with low computational time are normally used to try to reduce the training time of the model. On Table 1 you can see different examples of some of the most used activation functions nowadays.

Activation function	Mathematical formula	
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$	(Eq.1)
ReLU	$f(x) = \max(0,x) = \begin{Bmatrix} 0\ for\ x < 0 \\ x\ for\ x \geq 0 \end{Bmatrix}$	(Eq.2)
Leaky ReLU	$f(x)\max\ (ax,x) = \begin{Bmatrix} 0\ for\ x < 0 \\ a * x\ for\ x \geq 0 \end{Bmatrix}$	(Eq.3)
Softmax	$f(x) = \frac{e^{xi}}{\sum_{j=1}^{n} e^{xj}}$	(Eq.4)

Table 1. Common activation functions.

What is a layer?

Typically, neurons are aggregated into layers. The information goes through the layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (input layer) to the last layer (output layer), possibly and usually passing through multiple intermediate layers (hidden layers). A network is called a deep neural network if it has at least two hidden layers. On Figure 6 we can see a visual example of the different layers mentioned.

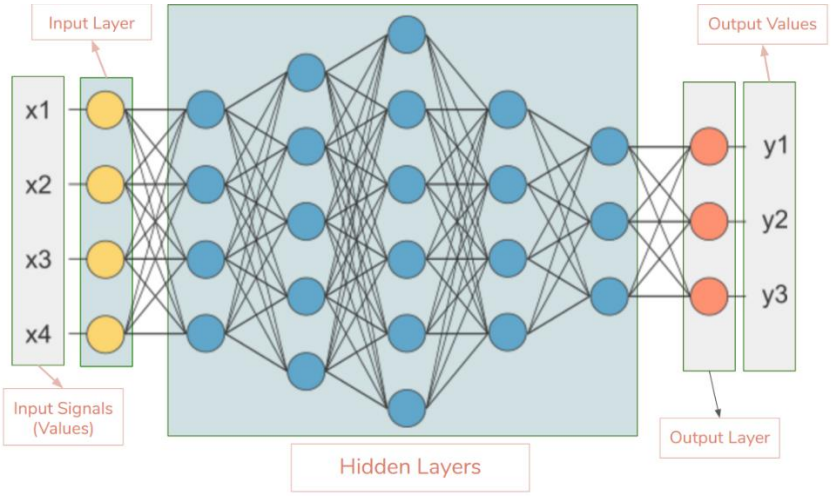

Figure 6. Example of the different layers found in a neural network.

Depending on the layers used, and how these are connected between them, we will obtain a different neural network architecture. The architecture of our neural network is a very important decision, we must select an adequate architecture for the task we want to realize.

In our case, for this project, we are going to work with CNNs. CNNs is a type of architecture specifically designed for processing grid-like data, such as images. They have revolutionized computer vision tasks and achieved state-of-the-art performance in various image-related-tasks, including processes like image classification, object detection, image segmentation, and more.

The key components of a CNN are convolutional layers, pooling layers and fully connected layers:

- **Convolutional layers:** Convolutional layers are the building blocks of a CNN. They consist of learnable filters or kernels that convolve across the input image to extract local features. Each filter performs a convolution operation by computing dot products between its weights and a small receptive field of the input. The outputs of multiple filters are stacked to create feature maps that highlight different visual patterns or features in the input.
- **Pooling layers:** Pooling layers help reduce the spatial dimensions of the feature maps while retaining the most relevant information. The pooling operation (typically max pooling, although there are others) downsamples the feature maps by selecting the maximum value within each local region. This downsampling helps make the representations more invariant to small translations and reduces the computational complexity of the network.
- **Fully connected layers:** Fully connected layers are typically placed at the end of the CNN architecture. They take the flattened feature maps from previous layers and

connect every neuron to every neuron in the subsequent layers. These layers help in making high-level predictions or classifications based on the learned representations.

During the training, CNNs use a technique called backpropagation to adjust the weights of the filters and fully connected layers based on the error between the predicted output and the ground truth labels. This optimization process aims to minimize the difference between predicted and actual outputs, improving the model's ability to make accurate predictions.

Quick explanation of how neural networks are trained

Training a neural network, or CNN involves different steps. Firstly, the model takes the input data and passes it through multiple layers to make a prediction. The difference between the prediction and the actual target/label is calculated using a loss function. The model then uses an optimizer to adjust its weights based on the gradients of the loss function, a process called backpropagation. This cycle repeats for multiple epochs until the model learns to make accurate predictions. There are multiple hyperparameters that we can modify in order to specify a different configuration for training our model. By modifying these parameters, we can adjust the model to the configuration that best suits the problem we are trying to solve.

Most used hyperparameters:

- **Epochs:** One epoch is a complete pass through the entire training dataset by the model during training. The number of epochs will determine how many passes the model does during an entire training session.
- **Batch size:** The batch size is the number of training examples used in one iteration to update the model's weights.
- **Loss function:** The loss function is a way to measure how well the model's predictions match the actual targets/labels. It's used to guide the model on its learning process.
- **Learning rate:** The value that determines the step size at which the model's weights are updated during training.
- **Optimizer:** An algorithm that adjusts the model's weights based on the gradients to minimize the loss function.

Gradients in machine learning are the partial derivatives of a loss function with respect to the model's parameters (weights). They indicate the direction and rate of change in the loss function as the parameters are adjusted. During training, gradients are used to update the model's weights in a way that minimizes the loss, guiding the model to improve its predictions.

2.2.3. Transfer learning

Transfer learning is referred to as a set of methods that allows us to transfer acquired knowledge from solving some problems to solve other problems. Transfer learning plays a very important role in deep learning. Normally, models used in this field need large amounts of time, data and resources to be trained. On the other hand, if we use these pre-trained models as a starting point, transfer learning allows us to develop quickly, efficient models and solve complex problems.

It mainly consists of building a new model on top of a prior one. These models are often trained on large-scale datasets like ImageNet (which contains over 14 million images as of today) and have learned to recognize a wide range of features and patterns. The goal is to enable this new model to leverage the knowledge from the pre-trained one, allowing the new model to be trained to its purpose faster and with less data. This technique is called fine-tuning. An example of transfer learning can be seen in Figure 7.

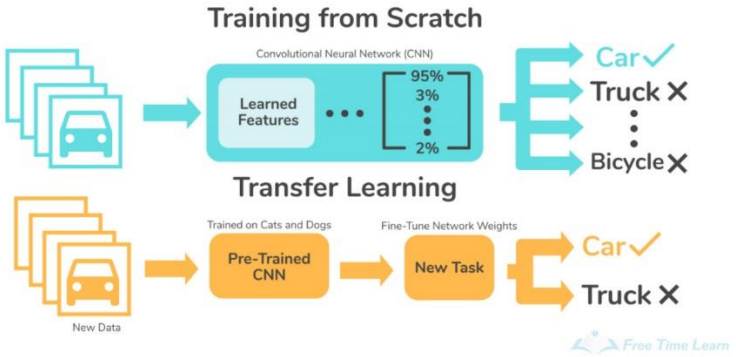

Figure 7. Example of training a neural network from scratch vs applying transfer learning.

2.2.4. Backbone of this project: EfficientNet B3

As previously explained, depending on how the layers are connected between them, we obtain a different architecture. Nowadays, one of the most used CNN architectures are EfficientNet models. EfficientNet models have demonstrated notable efficacy in achieving accurate results while managing computational efficiency. For this project, I will be using one of the EfficientNet models.

In Table 2 you can see a comparison between the different EfficientNet models. These models have been trained on the ImageNet dataset previously mentioned. Parameters, accuracy and GFLOPS can slightly vary depending on the library you use, as each one has its own implementation.

Model	Input Size	Parameter	GFLOPS	Accuracy (%)
EfficientNetB0	224x224	5.3M	0.39	93.532
EfficientNetB1	240x240	7.8M	0.69	94.186

EfficientNetB2	260x260	9.1M	1.09	95.31
EfficientNetB3	300x300	12.2M	1.83	96.054
EfficientNetB4	380x380	19.3M	4.39	96.594
EfficientNetB5	456x456	30.4M	10.27	96.628
EfficientNetB6	528x528	43.0M	19.07	96.916
EfficientNetB7	600x600	66.3M	37.75	96.908

Table 2. EfficienNet models comparison on ImageNet dataset (PyTorch library implementation).

The models of our interest for this project are the smaller versions of EfficientNet, specifically models B0, B1, B2 and B3. These models have fewer parameters but still have a significant performance in comparison with larger models that need more resources. Also, the smaller models play in our favor because they have been trained with smaller images. It is a common misconception that when using these pretrained CNN models, images need to be resized to the size of the images that were used to train these models. On the contrary, popular CNN are full convolutional networks that can accept any input size. The only relevance of the pretraining in smaller sizes is that these CNNs have learned to find certain patterns of certain sizes. This means that we do not need to use the exact same input sizes, but if we have small images, smaller models may perform better, and this is why we are focusing on these models in particular. The final model selected was EfficientNetB3. Mainly because it offers a great balance between quantity of parameters and a good performance. If we observe the Table 2, bigger models use a larger quantity of parameters and resources and do not outperform EfficientNetB3 by a lot. EfficientNetB3 architecture can be seen in Figure 8.

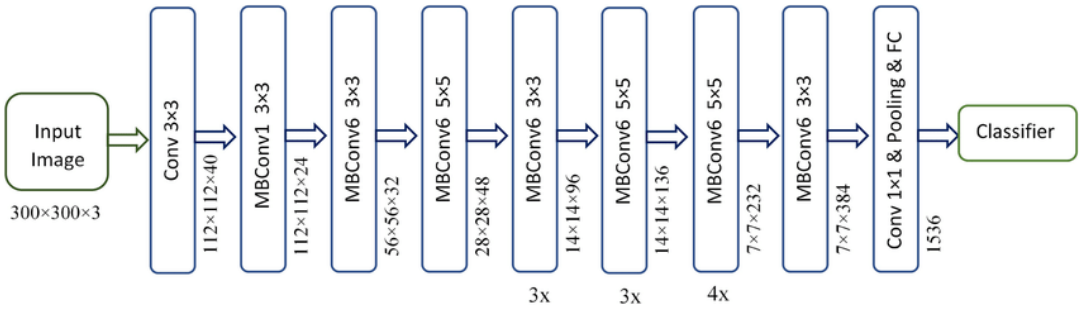

Figure 8. EfficientNetB3 architecture.

I will not make any major changes to the architecture. For the models with metadata, the metadata will go through some layers before being concatenated to the image predictions and finally go through the last layer. The model scheme can be seen in Figure 9.

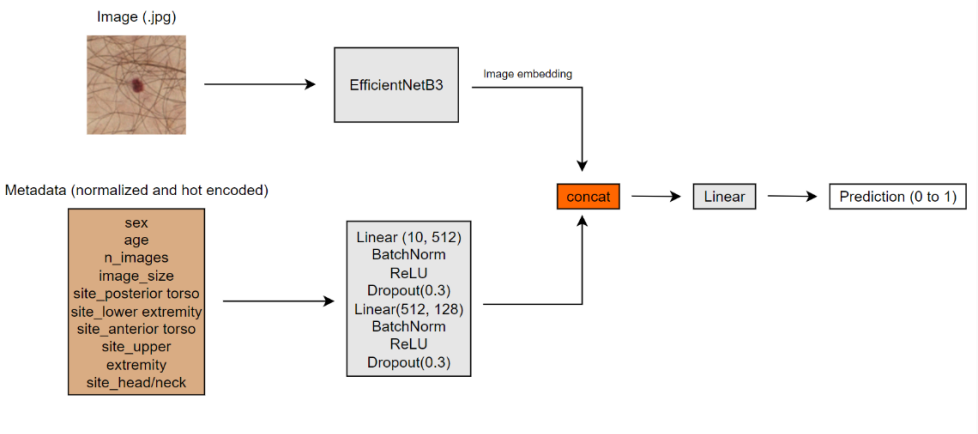

Figure 9. Own model implementation scheme.

2.2.5. Application of AI to skin cancer detection

As stated before, AI has gained a big role in multiple fields, including healthcare, in recent years. The application of AI in the field of dermatology, particularly in skin cancer detection, has gained a considerable attention, and here are some of the reasons why it is a good idea to use AI in skin cancer detection:

1. **Early and accurate detection:** AI, and more particularly machine learning and deep learning, have demonstrated the ability to analyze complex patterns in medical images. In our case, AI models can be trained to recognize small differences between benign and malignant lesions that could be overlooked by even experienced clinicians. Moreover, in this case, early detection of melanoma is crucial as it has a dramatic impact on improving the patient outcomes.

2. **Capacity of handling large volumes of data:** The increasing availability of image datasets, such as those from the ISIC challenges and the iToBoS project, among others, presents an opportunity to develop robust AI models.

3. **Reducing diagnostic variability:** Human dermatologists, despite their expertise, can have varying opinions on the same lesion. This leads to inconsistencies in diagnosis. An AI model, once trained, could provide consistent results, reducing this variability and ensuring a more standard approach.

4. **Helping in clinical decision-making:** AI systems can be used as decision-support tools to help dermatologists. The point is not to substitute the dermatologists, but to complement them by giving them a very powerful tool to assist them in making more informed decisions, ultimately leading to a better patient care.

5. **Efficiency and speed:** AI-powered systems can analyze images and data much faster than humans. This efficiency can be particularly valuable when resources are limited or the volume of patients exceeds the capacity of available dermatologists. In this case, AI could play a big role helping for example in discarding nor malignant cases or cases with a very low probability so the high-risk cases could receive the attention needed.

6. **Advances in AI techniques:** Recent advances in AI have shown remarkable success in image recognition tasks, so highly accurate melanoma detection models can be developed even with the limited annotated data available.
7. **Global impact and accessibility:** AI can have the potential to democratize healthcare by making expert-level diagnostic capabilities accessible in remote and underserved regions. A well-trained AI model can be deployed globally, allowing a general detection and early detection of melanoma, even in areas lacking specialized dermatological expertise.

As AI continues to evolve, its role in dermatology will expand. Additionally, AI-driven tools could be incorporated into teledermatology platforms, enabling patients to receive quick and accurate evaluations of their skin lesions from the comfort of their homes.

Chapter 3

State of the art

Nowadays skin lesion detection traditionally relies on manual methods performed by dermatologists. This includes visual, inspection, dermatoscopy (examination of the skin lesions with a dermatoscope), and biopsy in case the clinician is suspicious of any lesion to be malignant.

These methods are effective, but have many disadvantages. They are highly dependent on the clinician's experience and can be extremely time-consuming. We must consider that the average time spent analyzing a mole can be between 30 seconds and 2 minutes. This includes the visual inspection and the possible use of the dermatoscope to take an image of it. For someone with high number of moles this total inspection can sum up to more than 30 minutes (assuming that a very big quantity of the moles to revise has been discarded on the initial visual inspection, and only suspicious moles are checked one by one).

The application of this method has caused the accumulation of big dermatoscopic image collections of melanomas, other skin-cancer, and multiple types of skin-lesions images. These images have been collected in big datasets. Although there are multiple public smaller datasets used in different papers and research studies, the bigger datasets have been created or brought together mainly by ISIC (International Skin Imaging Collaboration). ISIC has used these datasets to sponsor annual challenges for the computer science community in association with leading computer vision conferences. Over the years, the challenges have grown in scale, complexity and participation, using high-quality human-validated training and tests sets of thousands of images and metadata. By 2018, the top-performing algorithms were consistently surpassing clinicians in diagnostic accuracy during "reader studies" (studies where the performance of medical professionals is compared to that of diagnostic tools or algorithms). Additional challenges in 2019 and 2020 were designed to address the out-of-distribution problem and assess the impact of clinical context respectively. Till this day ISIC has sponsored a total of 6 challenges including this year's 2024 challenge. These datasets are public and can be easily accessed through the ISIC Archive or through competitions platform's like Kaggle.

Luckily, during the last years, technology has made a big step in the world of dermatology and tools like the 3D TBP scanners have been created. 3D TBP (Total Body Photography) scanners are advanced imaging systems that allow us to create a detailed, three-dimensional map of a person's entire skin surface. These scanners are conformed by many cameras, that capture high-resolution images of the body from multiple angles. These images are combined to generate the 3D model of the skin. The main goal of these scanners is to allow the clinicians to monitor and track changes in skin lesions, moles and other dermatological features over time.

Figure 10. VECTRA WB360 3D TBP scanner by Canfield.

And this is where this project comes into play. These 3D TBP scanners, combined with AI models, can create a powerful tool together that can become the future of dermatology. And this is what project iToBoS (Intelligent Total Body Scanner for Early Detection of Melanoma) is all about.

3.1. Project iToBoS

iToBoS (Intelligent Total Body Scanner for Early Detection of Melanoma) project aims to develop an AI diagnostic platform for early detection of melanoma. The platform includes a novel total body scanner and a Computer Aided Diagnostics (CAD) tool to integrate various data sources such as medical records, genomics data and in vivo imaging of the patient skin surface.

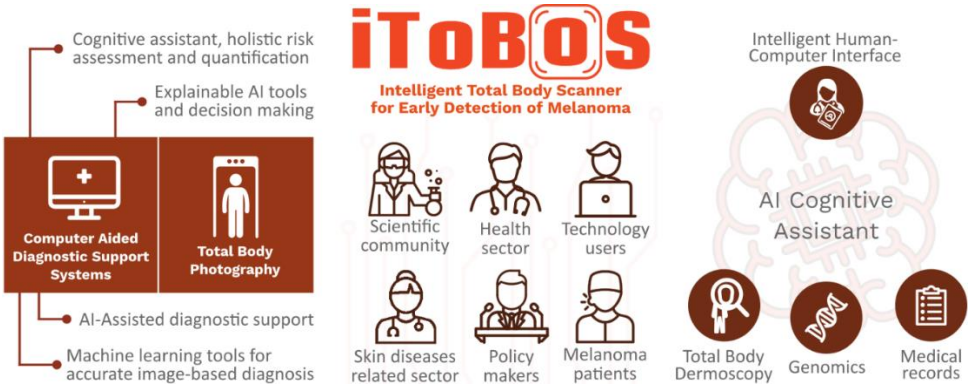

Figure 11. iToBoS project general scheme.

This approach will lead to a highly patient-tailored, early diagnosis of melanoma. The project will develop and validate an AI cognitive assistant tool to help healthcare practitioners, offering a risk assessment for every mole. Beyond integrating all available information about the patient

to personalize the diagnostic, it will provide methods for visualizing, explaining and interpreting AI model decisions to provide dermatologists valuable information for their clinical practice.

The new total body scanner will be based on an existing prototype developed by three of their project partners but powered with high-resolution cameras equipped with liquid lenses. These novel lenses will allow achieving unprecedented image quality of the whole body. The integration of such images with all available patient data using ML will lead to a new dermoscopic diagnostic tool providing prompt, reliable and highly personalized diagnostics for optimal judgment in clinical practice.

The project has a total of twenty participating partners, including powerful tech companies like IBM and Bosch and universities and hospitals from all around the globe. The University of Girona (UdG) is the general project coordinator. The project has a total budget of +13M €, and is led by Dr. Rafael Garcia Campos, the tutor of this project as well.

Chapter 4

Hypothesis and objectives

4.1. Research question

Can an artificial intelligence model be used as a powerful tool to service clinicians in skin-cancer early detection providing robust and confident results, by being able to predict correctly malignant and benign lesions?

4.2. Hypothesis

My hypothesis is that an artificial intelligence model can be used as a powerful tool to service clinicians in skin-cancer early detection providing robust and confident results, predicting correctly malignant and benign lesions.

4.3. Objective

The aim of this project is the development of an artificial intelligence model, based on a convolutional neural network, that has the mission of detecting malign skin-cancer from benign in crops images.

Chapter 5

Materials and methods

5.1. Regulations and legal aspects

Multiple models have been trained, and multiple datasets have been used during this project. In each dataset used, the images had a unique identifier for the image and a patient identifier.

In the case of ISIC datasets, because of being public anonymized data coming from hospitals and healthcare institutions, the data has already passed their own ethics committee.

In the case of the iToBoS datasets, the data coming from the University of Queensland (UQ) comes from the Brisbane Hospital and the data coming from Fundació Clínic per a la Recerca Biomèdica (FCRB) comes from the Barcelona Hospital Clinic, so both datasets have already been through their correspondent hospital ethics committee.

All the data used for this project was completely anonymous for the sake of the patients involved and their privacy.

5.2. Data

Multiple datasets have been used during this project. A total of two datasets have been used, one for training and one for test. As stated in previous chapters the data for this project consists of crops. These crops have been extracted from 3D TBP scanners.

For the training phase of the model, a dataset provided by the iToBoS project has been used. This dataset is composed by two separate datasets that have been combined. The datasets come from the University of Queensland (UQ) and the Fundació Clínic per a la Recerca Biomèdica (FCRB). An important reminder is that these crops are not from the iToBoS scanner that is currently being developed, these crops as said before come from 3D TBP scanners. Although these images have not been generated with the iToBoS scanner, they pertain to the iToBoS project as they are the images that may be used to create the AI assistant that will be installed on the scanner that is being developed. These two datasets combined sum up a total of 313571 images as seen in Table 3.

Dataset (training)	Benign	Malign	Total images
UQ	158	100017	100175
FCRB	186	213210	213396
	344	313227	313571

Table 3. UQ and FCRB datasets details.

The combination of both datasets leaves us with a total of 313571 available images, which will be divided into 80% for training and 20% for validation.

The test dataset will be the ISIC 2024 challenge dataset, which this year is about crops and not dermatoscopic images. Its details can be seen in Table 4.

Dataset (test)	Benign	Malign	Total images
ISIC 2024 challenge	393	400666	401059

Table 4. Test dataset details (ISIC 2024 trainig dataset).

As we can see, if we sum up both datasets, we have an enormous quantity of data available, a total of 714630 images. However, we have an important challenge, which is that these datasets are extremely unbalanced, with almost an average ratio of 1000:1. This means that for every 1000 benign images, we have 1 malignant, which is a big problem if we are trying to train an effective machine learning model. Such an imbalance can lead to a model that is biased towards predicting benign conditions, thereby reducing its ability to correctly identify malignant cases, which is crucial and even more in the medical context.

To try to address this issue, oversampling and undersampling techniques will be used to balance the training dataset during the training phase. I will use oversampling to increase the number of malignant cases by replicating them, while undersampling will be used to reduce the number of benign cases to try to achieve a more balanced dataset. In addition to this, extensive data augmentation process during the training of the models will be used too. This will contribute increasing the diversity of the training data, to help mitigate the effects of the imbalance. By using these strategies, I aim to make sure that the model is exposed to a more balanced dataset, where it can have a good representation of benign and malignant cases to improve its ability to predict correctly.

Oversampling and undersampling techniques are only used with the training portion of the training dataset. This means the validation and test dataset remain untouched. This is because validation and testing datasets are not submitted to data augmentation process, because it would not simulate real world situations where images have not gone under any process.

The best model obtained has been trained with an oversampling of 150 and an undersampling of 1:1 ratio. This means that for every malignant image of the training dataset, 150 artificial images have been created. The process followed can be seen in Table 5.

Dataset (original)	Benign	Malignant	Total Images
Training (80% of training dataset)	250581	275	250856
Validation (20% of training dataset)	62646	69	62715
Testing	400666	393	401059
Dataset (after OS & US)	**Benign**	**Malignant**	**Total Images**
Training 1:1 ratio	41250	41250	82500
Validation	62646	69	62715
Testing	400666	393	401059

Table 5. Datasets comparison and datasets used to train the best model.

A very big OS ratio is affordable because of the extensive data augmentation process I realize, that makes very unlikely to make two images to be the same.

5.3. Preprocessing

In this section, the preprocessing methods that have been used for the images will be explained. Some preprocessing methods that haven't been used will be explained as well, because I think that can be useful to know and are of interest for the project.

5.3.1. Data augmentation

Data augmentation is the process of artificially generating new data from existing data, primarily to train ML models. ML models require large and varied datasets, but sourcing sufficiently diverse and large real-world datasets can be challenging because of many limitations. Data augmentation techniques help enrich datasets by creating variations of existing data by applying certain augmentations. This provides the model a larger dataset for training and enables the model to see more diverse features. With this technique we help the model to better generalize to unseen data and improve its performance overall in real-world environments or cases. Data augmentation is essential as well to prevent overfitting and help to balance unbalanced datasets like this case.

For this project I will be using the Albumentations library, which allows an easy and quick integration with the PyTorch environment. A total of nineteen data augmentations will be used including two hair augmentations that will be explained more in depth in the next section.

Augmentations: Normalize, CustomHairAugmentation, CustomHairDrawingAugmentation, Transpose, VerticalFlip, HorizontalFlip, RandomBrightnessContrast, RandomBrightness, RandomContrast, MotionBlur, MedianBlur, GaussianBlur, GaussNoise, OpticalDistortion, GridDistortion, ElasticTransform, HueSaturationValue, ShiftScaleRotate, Resize and Cutout.

Figure 12 is an illustration of the before-after of these augmentations on four example images.

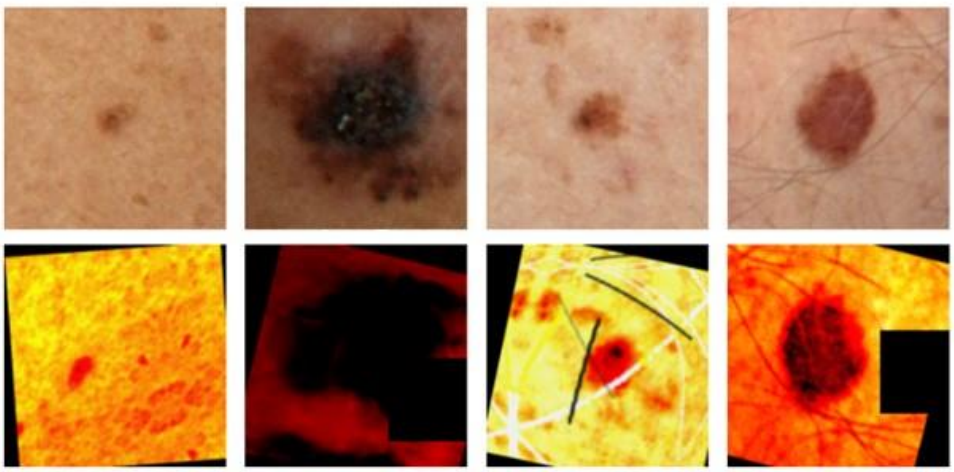

Fig 12. Training augmentation of four random images. Top row: original images; bottom row: augmented images.

5.3.2. Resizing of the images and comparison between crops and dermatoscopic images.

We must have in mind that these images are crops. So, they are images that have cropped from a 3D map of the skin patient that is no more than a bigger image. That means that in comparison to dermatoscopic images these crops have very low resolution, with most images being sized from around 150x150 to 250x250 (bigger and smaller images can be found as well). This means that we are going to be working with very low-resolution images and very small images, something that will difficult our task. In this case, different models ranging from 144x144 and 192x192 image size have been tested. Finally, almost all models have been trained with an image configuration of 192x192.

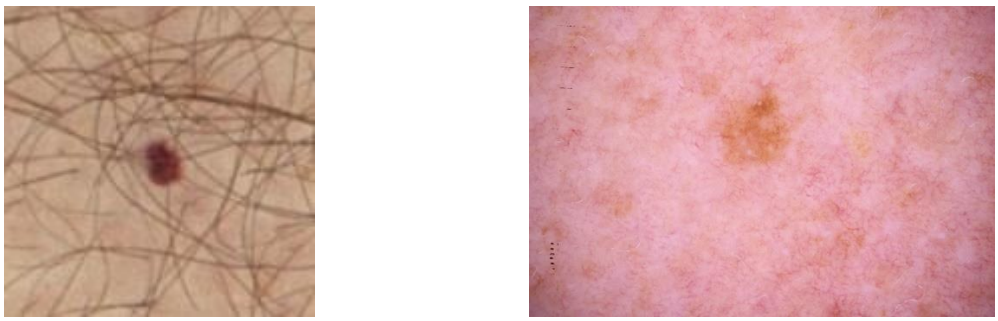

Figure 13. Comparison between two images to see the resolution difference. iToBoS to the left (155x155) and ISIC 2020 challenge to the right (6000x4000).

All images are resized before training the model, but more size usually means more detail in the image so more characteristics and patterns available to learn for the model.

Although machine learning models, particularly deep learning models, trained on high-resolution data tend to give better results, these models tend to perform worse when they are used on real-world scenarios, where the image quality may vary and will not be in the perfect conditions that have been seen during the training. So, although the low resolution hardens our task, it gives us the confidence that if a great model is obtained its performance will probably not vary if applied in the real-world.

Even so dermatoscopic images present other problems as well, for example artifacts and dark corner artifcats (DCA). Not all images have these issues but a big chunk of them do have atleast one of them.

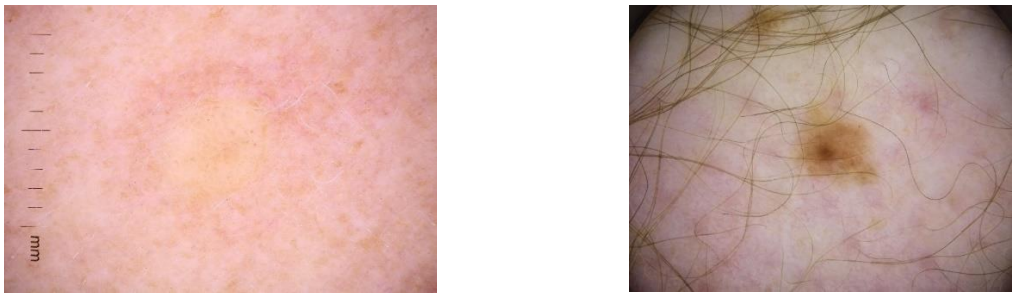

Fig 14. Artifacts (left) and DCA (right) in two images from ISIC 2020 challenge.

By artifacts we usually refer to the dermatoscop scale. Although the models can learn to ignore them, these artifacts can affect to the model performance. To this day research and novelty have not answered with confidence if these artifacts affect to the model performance in a good or bad way, but it's something to take into account if we are going to train a model with dermatoscopic images.

If wanted, there are some methods that try to tackle these artifacts.

Artifacts

- **Hair removal methods:** Hair removal approaches tend to delete as well these type of artifacts because they are similarly shaped to hairs. There are multiple existent methods, two of the most popular are DullRazor and blackhat algorithm. Nowadays the use of blackhat algorithm predominates. Although removing hair would let us think that the model performance would increase because it would not have any distractions, it actually isn't like that. In hair removal, the algorithm replaces hair pixels with combinations of the surrounding areas, which might not include relevant features and, in this way, affect the model negatively, erasing important information from the image.

Dark corner artifacts (DCA)

- **Smart fill-up algorithms:** Consists of using algorithms that fill up the DCA with pixel combinations of the surrounding areas, making it more realistic and similar to the skin.
- **Vignetting:** Apply a correction vignetting filter that compensates the darkness of the borders, regulating the brightness of the image.
- **Cropping:** This consists of cropping the image to not include these DCA. This can leave us with smaller images.

5.3.4. Hair augmentation

In comparison to hair removal, hair augmentation is actually helpful for the model. Although not significantly, I have seen a slight accuracy increase (1-2%) in the models that I have trained with hair augmentation in comparison to those without. This might be because hair augmentation provides a more realistic input to the model, resembling actual skin images more closely, as skin typically has some hair around or on it. In this project I have applied two different techniques of hair augmentation, artificial hair augmentation and realistic hair augmentation. These implementations are not of my own. They were created by other people in the ISIC 2020 Kaggle competition. What I have done is take them and transformed them into compatible albumentations modules. Albumentations is the library of data augmentation I am using for this project. Transforming this hair augmentation algorithms into compatible albumentations modules was a must to be able to use them correctly with the rest of augmentations. In the code, the sources from which these pieces of code have been extracted are provided.

The difference between them is very simple as their name indicates. The realistic hair augmentation consists of placing real hairs (that have been extracted from another image) on top of the image. And on the other hand, the artificial hair augmentation consists of drawing

artificial hairs on the image (mostly simple straight black lines). A comparison between both augmentations can be seen in Figures 15 and 16.

Figure 15. Realistic hair augmentation example.

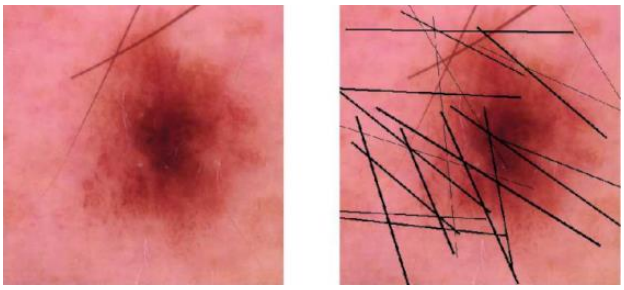

Figure 16. Artificial hair augmentation example.

5.3.5. Pacient metadata

In addition to the images of the lesions we have aswell metadata from the patients. All the metadata has been preprocessed. In particular I have used 10 metadata features in some of the models I have trained: sex that has been binarized, age_approx that has been normalized, n_images, image_size and 5 hot-encoded anatom_site_general features. This metadata go through two fully connected layers before being concatenated with the CNN features which then go to the final fully connected layer. The model architecture with metadata is illustrated in Figure 9 in Chapter 2.

5.4. Model training

The training script follows a remarkably simple pipeline, which can be easily read on the code files. In this section some of the decisions taken for the training of the model will be explained.

5.4.1. Optimizer

There are many optimizer algorithms. Common optimizers include Adam, SGD, SGD with momentum, RMSprop and AdaGrad, Adadelta and NAG among others. For this project, I decided to use Adam which is a popular optimization algorithm used in machine learning, particularly for training deep neural networks. Some of the reasons behind this choice:

- **Adjusts learning rate automatically:** Adam automatically adjusts the learning rate of each parameter, improving convergence speed and performance.

- **Efficient and fast:** Adam is very computationally efficient and well-suited for large datasets like this case.
- **Good default settings:** Adam typically runs well with default settings, reducing the need for an extensive hyperparameter tuning.
- **Handles noisy data:** Adam effectively manages sparse gradients and noisy data, common in deep learning.
- **Robustness and convergence:** Adam has shown to be robust and efficient in many deep learning applications.

5.4.2. Loss function

The loss function selected has been BCEWithLogitsLoss. However, some tests were realized with Focal Loss, a loss function that puts more emphasis on the hard to classify examples. The performance of it was not what I expected so I sticked with my original choice. Reasons behind the choice of this loss function:

- **Combination of sigmoid layer and BCE:** It combines sigmoid activation and binary cross-entropy loss into one function. This gives us numerical stability.
- **Direct use for binary classification:** Specifically designed for binary classification. It is ideal for our case.
- **Handling imbalanced datasets:** Allows different weights for classes, useful in imbalanced datasets like ours.
- **Probabilistic interpretation:** Outputs from BCEWithLogitsLoss can be directly interpreted as class probabilities.
- **Effective gradient computation:** Integrating Sigmoid with BCE ensures smoother, more reliable gradient computation and training.

5.4.2. Learning rate scheduler

A learning rate scheduler is a tool used in training neural networks that adjusts the learning rate during training. As the training goes on, it may be beneficial to reduce the learning rate to help the model converge more smoothly and avoid overshooting the optimal solution. The scheduler adjusts the learning rate according to a predefined scheduler or condition. In this case I am using ReduceLROnPlateau as scheduler provided by torch library. This scheduler will reduce the learning rate when a metric has stopped improving. I have used AUC as metric. The configuration used for the scheduler:

- **Patience = 3.** This means the model will wait 3 epochs without improvement on AUC before reducing the learning rate.
- **Factor = 0.05.** This is the factor by which our learning rate will be multiplied if it gets reduced.

- **Threshold = 0.01.** This is the threshold by which AUC needs to be improved in order to not make the learning rate scheduler start counting.

5.4.3. Early Stopping

Early stopping is a technique used to prevent overfitting during the training. It monitors the model's performance on the validation set and stops training when the performance stops improving, even if the training process has not completed the number of epochs predefined. By this way we save computational resources and prevent overfitting. This means that the model's weights will only be saved if the performance has improved. In my case I have used AUC as metric. An example, if we are at epoch number 2 and AUC is 0.8 the model will only be saved again if the AUC improves in the next epochs before finishing the training or early stopping. The configuration used:

- **Patience = 7.** This means the model will wait 7 epochs without improvement on AUC before stopping the training.

I am using an early stopping implementation from the library WTFML ("Well That's Fantastic Machine Learning).

5.4.5. Parameters and configurations

Multiple parameters and different configurations have been tested to acquire the best possible model. These are the modificable configuration parameters for the training phase.

- **Image size:** 144x144 and 192x192 are the sizes that have been tested. I finally decided to stick with 192x192 due to having a better performance, suiting with the original image sizes and being nearer the size of the images that have been used to train EfficientNetB3.
- **Batch size:** I have tried with batch sizes from 8-128. Finally decided to use an intermediate point 64.
- **Initial learning rate:** I have played with learning rates from 5e-4 to 1e-5. The best model acquired used 3e-5 as initial learning rate.
- **Pos weight:** Pos weight is a configuration parameter on the loss function BCEWithLogitsLoss that allows us to adjust the weights in the loss function. By using this what we do is that the FN have a bigger impact in the model optimization, improving this way the sensitivity of the model. It is a very useful tool in imbalanced datasets. I have tried values in the range of 2-10. The best model used a pos_weight = 10, this means that the errors in classifying the malignant class will have 10 times the impact in the optimization of the model.
- **Oversampling:** I have tried values in the range of 2-200. The higher the oversampling ratio the higher the sensitivity. Past 150, the metrics do not seem to improve.

- **Undersampling:** I have tried values in the range of 1-4. The best ratio for a great sensitivity results is 1:1. As we increase the ratio (we have more benign lesions than malignant), sensitivity drops significantly.
- **Use of metadata:** The use of metadata does not affect a lot to the performance, but it does not improve it either, more of this will be discussed in the results chapter.

Parameter	Value
Image size	192x192
Batch size	64
Initial lr	3e-5
Pos weight	10
Oversampling	150
Undersampling	1:1 ratio
Use of metadata	no

Table 6. Parameter configuration of the best trained model.

5.5. Metrics

A diversity of evaluation metrics exist to analyze AI models exist. In this project, multiple metrics have been used to track the model's performance during the training and test phases.

ROC-AUC-Score

Is a metric that measures the ability of a model to distinguish between classes. It is the area under the ROC curve (Receiver Operating Characteristic curve). The AUC value ranges from 0 to 1, with a higher value indicating a better model. The AUC is calculated as the integral of the ROC curve.

F1 Score

The F1 Score is the harmonic mean of precision and recall, balancing the two metrics.

Sensitivity

Sensitivity (True positive rate) measures the proportion of actual positives that are correctly identified by the model.

Accuracy Score

Accuracy is the proportion of correct predictions out of the total predictions made.

All these metrics mathematical formulas can be seen in Table 7.

Metric	Mathematical formula	
ROC-AUC Score	$AUC = \int_0^1 TPR(FPR)dFPR$	(Eq.5)
F1 Score	$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$	(Eq.6)
Sensitivity	$Sensitivity = \frac{TP}{TP + FN}$	(Eq.7)
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	(Eq.8)

Table 7. Metrics.

In the case of F1 Score and Accuracy, its balanced variants provided by sickit-learn library have been used because of the imbalance of the dataset to provide realistic results.

ROC Curve

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. It is a plot of the true positive rate (Sensitivity) against the false positive rate (1 - Specificity).

Confusion matrix

A confusion matrix is a table used to evaluate the performance of a classification model by comparing the actual target values with those predicted by the model. It provides the counts of true positives, false positives, true negatives, and false negatives. The confusion matrix is a valuable tool for understanding the performance of a classification model in detail, particularly when dealing with imbalanced datasets. It helps you to visualize not only how often the model is correct but also the types of errors it makes. A scheme can be seen in Figure 17.

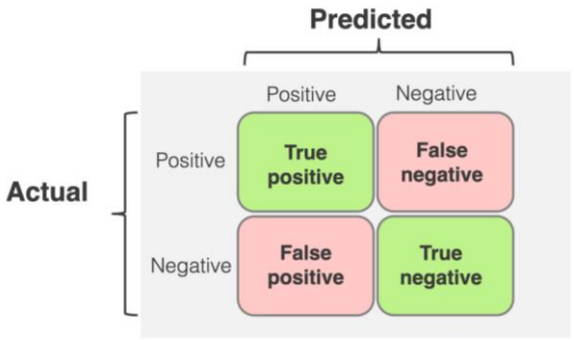

Figure 17. Confusion matrix scheme.

Chapter 6

Results

6.1. Model results and discussion

After training tens of models and trying a lot of things I accomplished to train a model that has in my point of view a great performance and has shown good results.

These are the results of the best model acquired:

iToBos_35e_192_64bs_3e-05lr_BCEWithLogitsLoss_OSx150_USRatio_1:1 on test set	
ROC-AUC Score	0.9107
F1 Score with a predefined threshold of 0.9	0.9028
Sensitivity with a predefined threshold of 0.9	0.8422
Accuracy Score	0.8333

Table 8. Best model results.

In the following figure you can see the ROC Curve on the test set and the confusion matrix.

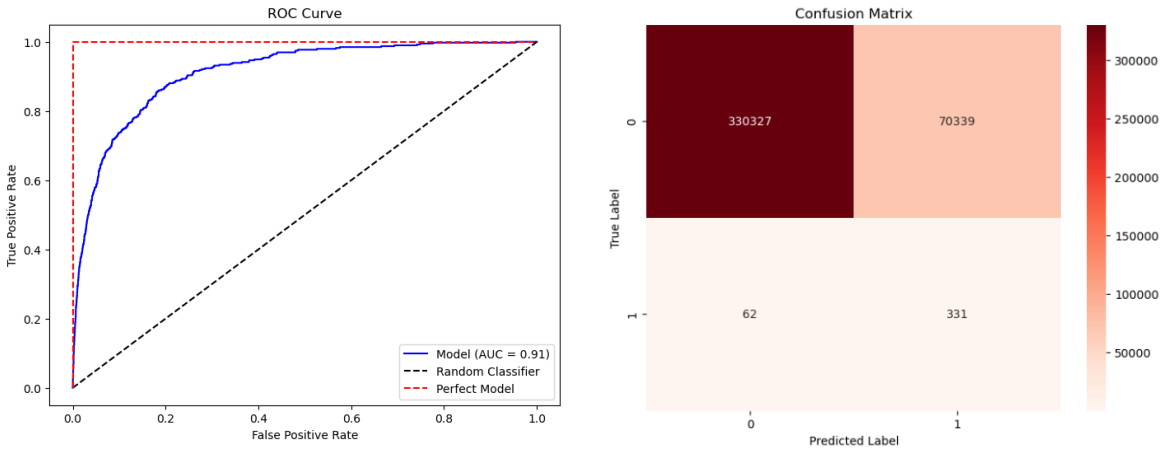

Figure 18. ROC Curve and confusion matrix of the best model.

From my point of view, these results are quite impressive considering that we are dealing with very low-resolution images. Results that I thought were not achievable with these kind of images at the start of this project. However, these results have room for improvement.

A higher sensitivity has been achieved in other models, around 0.94-0.95 with the same threshold of 0.9, but at the cost of generating a extremely big quantity of false positives. The only difference is that the models were using sensitivity as metric instead of AUC in the learning rate scheduler and early stopping algorithm. I analyzed and thought about it very careful but I preferred to stay with this model because of the high cost of improving the sensitivity. All the other metrics dropped significantly to around 0.6-0.7 and as mentioned before an extremely high number of false positives was generated. By extremely high number I am talking that for every TP I was gaining I was generating around five thousand false positives something that

from my point of view is not viable to apply to a real-world scenario. If the magnitude order was lower, and the cost of every TP was in the tens or even hundreds my thoughts would have been different, but we are talking of multiple thousands of false positives for every true positive.

Another way of improving the sensitivity would have been to drop the imposed threshold. This way better results would have been achieved but I thought that was a way to lie to myself about the results. Also, I decided that in the medical context to diagnose something with less than a 90% confidence was not viable.

The same model was trained with metadata to see if metadata was able to improve the performance. Other models were trained with metadata as well to see the performance difference. There was no big performance difference in most of the cases, and in almost all of them the sensitivity dropped slightly (around 5%). This case is not an exception as you can see in Table 9.

iToBos_35e_192_64bs_3e-05lr_BCEWithLogitsLoss_OSx150_USRatio_1:1 with metadata on test set	
AUC	0.9138
F1 Score with a predefined threshold of 0.9	0.9323
Sensitivity with a predefined threshold of 0.9	0.7913
Accuracy Score	0.8331

Table 10. Model with metadata results.

As you can see there is a slight increase on the AUC and F1 score but a slight drop as well on the sensitivity. In my opinion it's better to prioritize sensitivity because of the medical context so I will stick with the model that has been trained without metadata. In the following figure you can see the ROC Curve on the test set and the confusion matrix of this second model.

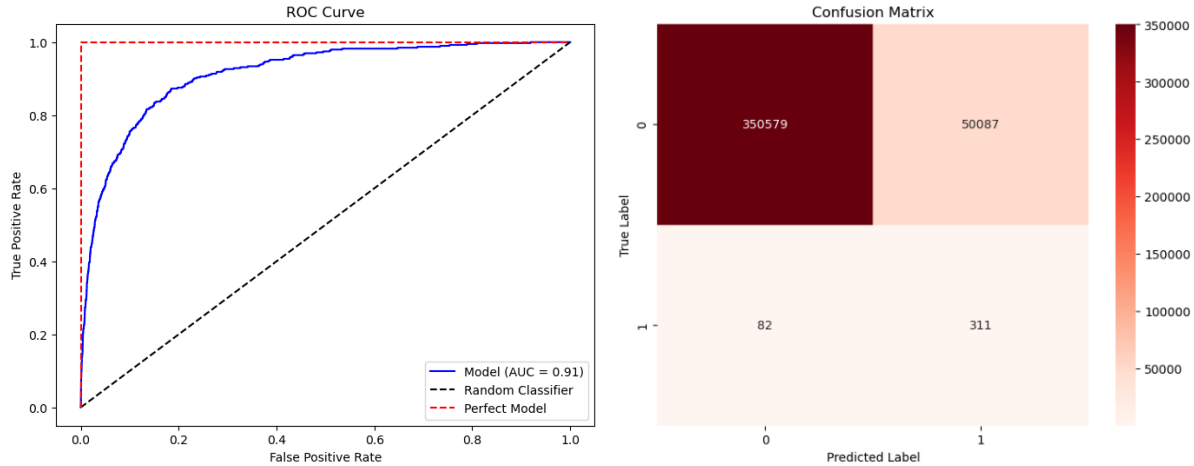

Figure 19. ROC Curve and confusion matrix for the metadata model.

As we can see there is no much difference between the two models other than the sensitivity and the changes its decrease suppose (more false negatives and less false positives).

If I had to choose between the use or not of metadata, now a days I would prefer to not use it as in this case seems to not increase the performance the way we would like. The use of metadata is a little bit tricky. From my point of view metadata models are a good idea in ensembles, there they can be useful because they contribute diversity.

I say the use of metadata is a little bit tricky because from what I have been working on I have not seen any patient metadata that increases the model performance. The genetic risk factor might be a useful data that could actually help to increase the model performance, but other than that metadata seems to not help a lot. Mostly because the model does not find any correlation between this data with a lesion being malignant or benign. This is why I say genetic risk factor could actually increase the performance, because is a data that can provide correlation.

Also the acquisition of metadata is very hard. The acquisition of metadata relies on clinicians and people that work on the hospitals. On processes where humans are involved there will be always errors, and this is no exception. The process sometimes is a form or survey where the patient writes it's data, sometimes the patient does not fill all the questions, sometimes the clinician does not remember to give the form, etc. This leads to NULL or empty values, different data formats in the same column, different ways of saying the same thing, there is no predefined standard. All these problems lead to not very high-quality metadata.

If a protocol and common standard was created for the acquisition of metadata, then my opinion in respect to it would change. But this is extremely difficult to implement in the real world. That's why only image models are more reliable and are easier to integrate in real-world scenarios, because they only depend on the image and its target/label.

Chapter 7

Discussion

7.1. Limitations

The realization of the project has been very satisfying and successful. However, I have found myself with some limitations and problems during these months. I would say I have had two main limitations:

- The first limitation would be the data, mainly the imbalance of the datasets. Although the problem can be somehow fixed with the methods we have used, the quality of the artificially generated data can not be compared to real data. Much better models could be created with big balanced and robust datasets. However it's completely understandable the imbalance of the datasets because of the number of benign lesions in comparison to the malignant lesions in humans skin.
- The second limitation would be time because of many reasons. The first of them is that PyTorch frame it's quite challenging for beginners and was something relatively new for me, so it took quite a lot of time for me to understand and get to know it at a decent level. And the second and obvious is that model training is an extremely high time consuming task. Having to wait multiple hours to see if the changes you have made have great results or not is very stressful and difficult to manage when you have a submission deadline. However, some things have played in my favor to reduce the model training time, like image size and the fact that I have not used cross validation for this project. The implementation of cross validation would have multiplied by five the model training time, making it easily tens of hours. Luckily as well I have had the opportunity to use ViCOROB servers and GPU that have made this training sessions much faster, if not this project would have impossible.

Even though I've had these limitations, the results obtained from this project have been really great.

7.2. Contributions to ODS

The main goal of this project is to develop an AI model that is able to help clinicians in their jobs. So, one of the objectives of this project is the third ODS, to ensure the achievement of a healthy life and promote well-being for people of all ages. The proper implementation of this project could help clinicians in their job to early detect melanoma and skin-cancer more accurately, opening the possibility to an earlier diagnosis and treatment which would potentially save more lives and improve patient lives. This project could also be fitted in the ninth ODS, industry, innovation, and infrastructure. Mostly because we are aiming to develop and implement advanced AI technology in the healthcare field. So, this project promotes the technology innovation and infrastructure improvement in the healthcare system.

Fig 20. 3rd ODS

Fig 21. 9th ODS

Chapter 8

Conclusions

To conclude, this project successfully accomplished the primary objective of creating a robust and precise AI model that can be used as a tool by clinicians. The model accomplished has reached a ROC-AUC of 0.91, an F1-score of 0.90, a sensitivity of 0.84 (with a predefined threshold of 90%), and an accuracy of 0.83. A precise and robust model able to differentiate and predict correctly benign and malignant lesions has been developed. We've been able to see how AI can, once implemented in the medical field be a very powerful tool, not only in the early diagnosis of skin cancer and melanoma but among all the other fields of healthcare.

This project demonstrates that a good AI model can be created from crops and that high resolution images as dermoscopic images are not strictly necessary to develop a robust model. It also confirms that once the iToBoS project scanners are finished and implemented with their AI assistant, they will be state-of-the-art technology in the medical field and will deliver an exceedingly high quality and tailored experience to the patient.

The development of this project has allowed me to expand my knowledge in AI and ML methods applied to healthcare. It has helped me to learn much more than what I could have imagined. And more importantly it may have showed me the path I want to continue in my professional career as a biomedical engineer.

In conclusion, the developed model shows us the enormous potential AI as a powerful tool to improve healthcare.

8.1. Possible improvements

Some of the possible improvements for this project:

- **Different ML methods ensemble:** At the same time as this report was being written, the ISIC 2024 challenge was taking place. To see if there was anything of interest for this project, I decided to keep an eye on it from time to time and found something quite interesting. To this day, different ML methods other than CNNs are predominating the challenge. One of the reasons is because a lot of people believe these crops do not have enough resolution to achieve a great performance with CNNs. The other reason is tabular data. These 3D TBP scanners apart of the image and the metadata that we have talked about in this project, also provide tabular data. This tabular data are different calculous and coordinates/positions of the lesion provided and done by the scanner. There are a lot of participants using methods like LightGBM, a ML method that is a

gradient boosting framework that uses tree-based learning algorithms. Most of the people that are using these methods seem to be getting good results due to LightGBM being very good and efficient for tabular data. Knowing all this, one of the possible improvements I propose would be an ensemble of an image model like the one that I have developed in this project and a tabular data model like this. By making an ensemble of these two models we would be using all the data available and would have a very diversified model that could probably outperform the model developed in this project. Although as I have mentioned before, using more data than the image can have its problems and inconveniences, in this case they are more unlikely given that this data is calculated and created by the scanner. In this way the possible human error disappears.

- **Multiple CNNs ensemble:** Another improvement would be the ensemble of multiple CNNs models and architectures with and without metadata to provide diversity.
- **A more extensive hyperparameter tuning:** Although unlikely, a more extensive hyperparameter tuning could improve the model results by a small percentage. However, the architecture and the data give the model a limit that cannot be surpassed.
- **Multiclass classification instead of binary classification:** Another improvement could be to build a model that classifies into classes instead of classifying in benign or malign lesions. This way, we could see directly how good the model is in classifying melanoma specifically and have metrics like melanoma sensitivity. However, multiclass models are much bigger and consume much more resources, and I am pretty sure that would not be as viable with such a low-resolution images as binary classification models are.

Chapter 9

References

1. American Cancer Society. What Is Melanoma Skin Cancer? | What Is Melanoma? [Internet]. www.cancer.org. 2019. Available from: https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/what-is-melanoma.html
2. Melanoma - Síntomas y causas - Mayo Clinic [Internet]. www.mayoclinic.org. Available from: https://www.mayoclinic.org/es/diseases-conditions/melanoma/symptoms-causes/syc-20374884
3. Cancer Research UK. Melanoma skin cancer | Cancer Research UK [Internet]. Cancerresearchuk.org. 2016. Available from: https://www.cancerresearchuk.org/about-cancer/melanoma
4. The Skin Cancer Foundation. Melanoma [Internet]. The Skin Cancer Foundation. 2018. Available from: https://www.skincancer.org/skin-cancer-information/melanoma/
5. Lopes J, Rodrigues CMP, Gaspar MM, Reis CP. Melanoma Management: From Epidemiology to Treatment and Latest Advances. Cancers [Internet]. 2022 Jan 1;14(19):4652. Available from: https://www.mdpi.com/2072-6694/14/19/4652#
6. Stryker C, Kavlakoglu E. What is artificial intelligence (AI)? [Internet]. IBM. 2024. Available from: https://www.ibm.com/topics/artificial-intelligence
7. What Is Artificial Intelligence (AI)? [Internet]. Google Cloud. Available from: https://cloud.google.com/learn/what-is-artificial-intelligence?hl=en
8. Wei ML, Tada M, So A, Torres R. Artificial intelligence and skin cancer. Frontiers in medicine [Internet]. 2024 Mar 19;11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10985205/
9. Beltrami EJ, Brown AC, Salmon PJM, Leffell DJ, Ko JM, Grant-Kels JM. Artificial intelligence in the detection of skin cancer. Journal of the American Academy of Dermatology. 2022 Aug;
10. Stanford University CS231n: Deep Learning for Computer Vision [Internet]. cs231n.stanford.edu. Available from: https://cs231n.stanford.edu/
11. Protocol R. Everything you need to know about Neural Networks [Internet]. Medium. 2018. Available from: https://medium.com/ravenprotocol/everything-you-need-to-know-about-neural-networks-6fcc7a15cb4

12. What is Transfer Learning? [Internet]. skyengine.ai. Available from: https://skyengine.ai/se/skyengine-blog/128-what-is-transfer-learning

13. Amad. Unleashing the Power of Transfer Learning in Artificial Intelligence [Internet]. Medium. 2023. Available from: https://iamamadsiddiqui.medium.com/unleashing-the-power-of-transfer-learning-in-artificial-intelligence-e8a9deee62f3

14. Mwiti D. Transfer Learning Guide: A Practical Tutorial With Examples for Images and Text in Keras - neptune.ai [Internet]. neptune.ai. 2021. Available from: https://neptune.ai/blog/transfer-learning-guide-examples-for-images-and-text-in-keras

15. M. Roshni Thanka, E. Bijolin Edwin, Ebenezer V, K. Martin Sagayam, B. Jayakeshav Reddy, Hatıra Günerhan, et al. A hybrid approach for melanoma classification using ensemble machine learning techniques with deep transfer learning. Computer methods and programs in biomedicine update. 2023 Jan 1;3:100103–3.

16. Kaur R, GholamHosseini H, Sinha R, Lindén M. Melanoma Classification Using a Novel Deep Convolutional Neural Network with Dermoscopic Images. Sensors. 2022 Feb 2;22(3):1134.

17. Ha Q, Liu B, Liu F. Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge [Internet]. arXiv.org. 2020 [cited 2024 Sep 3]. Available from: http://arxiv.org/abs/2010.05351

18. González-Cruz C, Jofre MA, Podlipnik S, Combalia M, Gareau D, Gamboa M, et al. Machine Learning in Melanoma Diagnosis. Limitations About to be Overcome. Actas Dermo-Sifiliográficas (English Edition) [Internet]. 2020 May 1 [cited 2022 May 6];111(4):313–6. Available from: https://www.actasdermo.org/en-machine-learning-in-melanoma-diagnosis--articulo-S1578219020300846

19. C. Kaushik Viknesh, Kumar P, Seetharaman R, D. Anitha. Detection and Classification of Melanoma Skin Cancer Using Image Processing Technique. Diagnostics. 2023 Oct 26;13(21):3313–3.

20. Samuel William Pewton, Moi Hoon Yap. Dark Corner on Skin Lesion Image Dataset: Does it matter? 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022 Jun 1;

21. Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Felix, et al. Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification. JDDG Journal der Deutschen Dermatologischen Gesellschaft. 2021 May 10;19(6):842–50.

22. Pewton SW, Cassidy B, Kendrick C, Yap MH. Dermoscopic dark corner artifacts removal: Friend or foe? Computer Methods and Programs in Biomedicine [Internet]. 2024 Feb 1 [cited 2024 Jul 23];244:107986. Available from: https://pubmed.ncbi.nlm.nih.gov/38157827/

23. Winkler JK, Sies K, Fink C, Toberer F, Enk A, Abassi MS, et al. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. European Journal of Cancer. 2021 Mar;145:146–54.

24. Naseer I, Akram S, Masood T, Jaffar A, Khan MA, Mosavi A. Performance Analysis of State-of-the-Art CNN Architectures for LUNA16. Sensors [Internet]. 2022 Jan 1 [cited 2023 Jan 3];22(12):4426. Available from: https://www.mdpi.com/1424-8220/22/12/4426

25. torchvision.models — Torchvision 0.10.0 documentation [Internet]. pytorch.org. Available from: https://pytorch.org/vision/stable/models.html

26. Masood D. Pre-trained CNN architectures designs, performance analysis and comparison [Internet]. Medium. 2023. Available from: https://medium.com/@daniyalmasoodai/pre-train-cnn-architectures-designs-performance-analysis-and-comparison-802228a5ce92

27. Sarkar A. Understanding EfficientNet — The most powerful CNN architecture [Internet]. Medium. 2021. Available from: https://arjun-sarkar786.medium.com/understanding-efficientnet-the-most-powerful-cnn-architecture-eaeb40386fad

28. Sovit Ranjan Rath. Transfer Learning using EfficientNet PyTorch [Internet]. DebuggerCafe. 2022 [cited 2024 Sep 3]. Available from: https://debuggercafe.com/transfer-learning-using-efficientnet-pytorch/

29. Primiero CA, Betz-Stablein B, Ascott N, D'Alessandro B, Gaborit S, Fricker P, et al. A protocol for annotation of total body photography for machine learning to analyze skin phenotype and lesion classification. Frontiers in Medicine [Internet]. 2024 Apr 9 [cited 2024 Jun 8];11:1380984. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11035726/

30. You K, Long M, Wang J, Jordan MI. How Does Learning Rate Decay Help Modern Neural Networks? [Internet]. arXiv.org. 2019 [cited 2024 Sep 3]. Available from: http://arxiv.org/abs/1908.01878

31. Shah S, Shah M, Pandya AS, Rajat Sushra, Ratnam Sushra, Mehta MP, et al. A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN). Clinical eHealth. 2023 Dec 1;6:76–84.

32. Pla B. Application of an Image Segmentation Method for Intracerebral Hemorrhage Images. Handlenet [Internet]. 2023 [cited 2024 Sep 3]; Available from: http://hdl.handle.net/10256/24130

33. Primiero CA, McInerney-Leo AM, Betz-Stablein B, Whiteman DC, Gordon LG, Caffery LJ, et al. Evaluation of the efficacy of 3D total-body photography with sequential digital dermoscopy in a high-risk melanoma cohort: protocol for a randomised controlled trial. BMJ

Open [Internet]. 2019 Nov 1 [cited 2023 Apr 27];9(11):e032969–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6858160/

34. Urbano B. Segmentació de lesions cerebrals procedents d'imatges mèdiques per a la seva posterior digitalització i impressió 3D. Handlenet [Internet]. 2022 [cited 2024 Sep 3]; Available from: http://hdl.handle.net/10256/21944

35. CORDIS cordis.europa.eu. Intelligent Total Body Scanner for Early Detection of Melanoma [Internet]. CORDIS | European Commission. 2024 [cited 2024 Sep 3]. Available from: https://cordis.europa.eu/project/id/965221/reporting/es

36. Ghadah Alwakid, Gouda W, Humayun M, Jhanjhi NZ. Diagnosing Melanomas in Dermoscopy Images Using Deep Learning. Diagnostics. 2023 May 22;13(10):1815–5.

37. iToBoS [Internet]. Trilateral Research. 2024 [cited 2024 Sep 3]. Available from: https://trilateralresearch.com/work/itobos

38. Nazari S, Garcia R. Automatic Skin Cancer Detection Using Clinical Images: A Comprehensive Review. Life. 2023 Oct 26;13(11):2123–3.

39. What is Data Augmentation? - Data Augmentation Techniques Explained - AWS [Internet]. Amazon Web Services, Inc. Available from: https://aws.amazon.com/what-is/data-augmentation/#:~:text=Data%20augmentation%20is%20the%20process

ANNEX A

Project planification

Step 1. Project set-up and research.

Duration: 1 month

In this part I started to search for information about the topic of the project.

Step 2. Learn PyTorch basics.

Duration: 1 month

During this part I started to learn and play with the PyTorch library. I coded my first scripts and started to train very small models with very few images in order to get used to the framework.

Step 3. Coding the pipeline.

Duration: 2 months

After having learned the basics, I started to code the different scripts needed to train the model. However, I continued learning PyTorch during the whole project duration. The more I learned, the more I realized I did not know anything and still had so much to learn. This part took me quite a long time because one of the first pipelines I coded was giving me extremely bad results for some reason, so I decided to build a new one from scratch, testing everything, so I made sure everything worked correctly.

Step 4. Model training.

Duration: 3 months

During this part, I started to train my first complex models. At first, I started with the ISIC dermatoscopic datasets in order to test the pipeline performance and to see the results. Then I finally started training models using the iToBoS datasets.

Step 5. Memory writing.

Duration: 1 month

Although some things were written all along the project and most of the research and sources of information were founded at the beginning, the memory was written on the last month.

GANTT Diagram

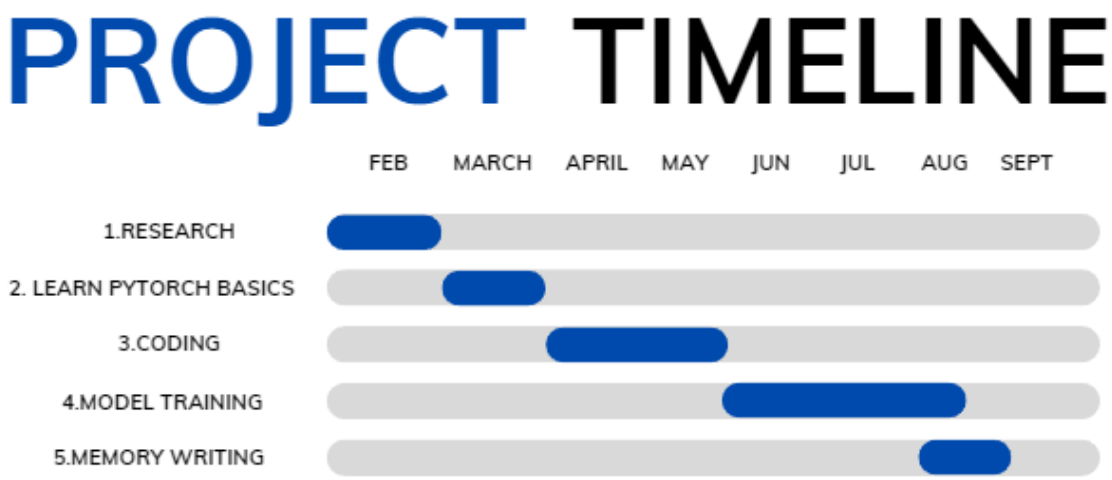

Figure 22. GANTT diagram of the project timeline

ANNEX B

Code

Following the required specifications, given that this project is carried out in software development, the code used is available on a GitHub repository to make it available to the scientific community or anyone who wants to check it out.

The repository has all the necessary things to reproduce the project, except the iToBoS data used for the training that to this day is not publicly available but might be in the future. However, the code can be used to train models using any of the ISIC datasets.

Link to the repository: https://github.com/ivgogo/tfg

ANNEX C

Budget

In this annex I will show the estimated costs of carrying out this project. The resources provided by ViCOROB (access to the server with the GPU to train the models) have been included in the budget.

C.1. Working Hours

Description	Hours	Price per hour (≈)	Cost
AI engineer	360	40€	14400€

Table 11. Working hours invested in the project.

C.2. Resources

Description	Quantity	Cost
NVIDIA A100 80 GB PCIe	1	19818.41€
ThinkPad e490 Laptop	1	980€
Microsoft Word License	1	6€

Table 12. Resources used.

C.3. Total budget

Description	Quantity	Hours	Price per hour (≈)	Cost
NVIDIA A100 80 GB PCIe	1	-	-	19800€
ThinkPad e490 Laptop	1	-	-	980€
Microsoft Word License	1	-	-	6€
AI engineer	1	360	40€	14400€
			Subtotal	35186€
			Taxes	21.00%
			Total pre-taxes	**35186€**
			Total with taxes	**42575,06€**

Table 13. Total budget of the project.

I want to emphasize again that this work would not have been possible without the resources provided by ViCOROB.

ANNEX D

Ethics committee

As previously mentioned, the data that has been used for this project is completely anonymized and has passed through its own ethics committee as most of it comes from hospitals or healthcare institutions. That said, there is no need of any ethical committee to approve this project because during all the development of it the rights, security and welfare of the patients are secured.

ANNEX E

Software and technologies

This project has been developed in Python. A high-level programming language, that is used in the development of a high quantity of applications. It is one of the most languages used in AI, Big Data, ML and Data Science. Its similarity to human language, makes it a very high understandable language for someone with basic programming skills. It has a high variety of libraries that lets developers their imagination to be the limit.

The framework selected for designing and training neural networks has been PyTorch. Although it has a harder entry barrier for beginners, it allows to create and test different experiments and pipelines very fast.

The project has been developed using Visual Studio Code (VSCode) as code editor. It allowed me to remotely connect to ViCOROB servers and to have an IDE (integrated development environment) where I had access to their powerful GPU and all the iToBoS project data. Linux (Ubuntu LTS 22.04) has been used as OS to be able to access easily to the server via terminal and because it's powerful utility for developing.