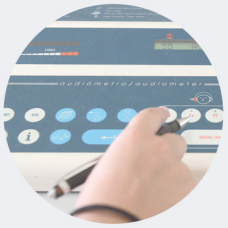
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA BIOMÉDICA






Determinação de Parâmetros de Risco no Melanoma com Técnicas de Deep Learning

DIOGO MIGUEL SOUSA PINTO
Novembro de 2022

Determinação de Parâmetros de Risco no Melanoma com Técnicas de Deep Learning

Diogo Miguel Sousa Pinto

Engenheiro(a) Biomédico(a) pelo Instituto Superior de Engenharia do Porto

"Dissertação apresentada no Instituto Superior de Engenharia do Porto para a obtenção de grau de Mestre em Engenharia Biomédica"

Orientador: Luís Coelho

Carlos Ferreira

[Outubro 2022]

Resumo

Em 2020, estima-se que 325.000 pessoas tenham sido diagnosticadas com melanoma, estimando-se que 50.000 mortes em todo o mundo por doença. Além disso, o melanoma é um dos cancros mais comuns em jovens adultos. O melanoma, quando em fase inicial, tem uma grande taxa de sobrevivência, com cerca de 97%, no entanto, em fases tardias a taxa de sobrevivência é apenas de cerca de 10%, o que implica que uma deteção precoce é extremamente importante.

Nos últimos anos, o machine learning tem feito progressos notáveis e com isso seguiu-se o número de pesquisas e utilizações em questões médicas. Seria possível utilizar os grandes avanços na inteligência artificial e na classificação de imagem para criar uma aplicação para detetar precocemente o melanoma?

Nesta tese são testadas três arquiteturas de machine learning para prever melanoma, VGG16, VGG19 e MobileNet V2. Todas as três arquiteturas usadas foram pré-treinadas e aplicadas a um dataset que compila todos os principais datasets de melanoma públicos. A arquitetura com melhores resultados foi o MobileNet V2, que com afinação alcançou mais de 96% de precisão. Foi também testado para aplicar o pré-processamento nas imagens do dataset (filtro CLAHE), que acabou por ter uma precisão ligeiramente menor na mesma arquitetura, com menos 0,2%. Com o melhor modelo, foi criada uma aplicação simples para permitir aos pacientes utilizarem o modelo de machine learning.

O modelo de aprendizagem automática criado nesta tese alcançou uma precisão excecionalmente elevada, em comparação com a literatura. Além disso, destaca-se por formar o resto porque utiliza uma arquitetura recente e leve e tem uma aplicação em execução na parte frontal que permite a todos, mesmo pessoas sem conhecimento técnico, utilizar a aplicação.

Palavra- chave: melanoma; inteligência artificial; machine learning; classificação de imagens; segmentação de imagens; supervised learning ; imagem médica; MobileNet V2; Python.

Abstract

In 2020, there were an estimated 325,000 people diagnosed with melanoma, and an estimated 50,000 worldwide deaths from the disease. Moreover, melanoma it's one of the most common cancers in young adults. Melanoma, when in early stages, has a great survival rate, with around 97%, however, in late stages the survival rate is only around 10%, which implies that an early detection is extremely important.

In recent years, deep learning has made outstanding breakthroughs and with that the number of researches and uses in medical issues followed. Would it be possible to use the great advances in artificial intelligence and image classification to create an application to early detect melanoma?

In this thesis three machine learning architectures are tested to predict melanoma, VGG16, VGG19 and MobileNet V2. All three architectures used were pre-trained and applied to one dataset that compiles all major public melanoma datasets. The architecture with best results was the MobileNet V2, that with fine-tuning achieved more than 96% of accuracy. It was also tested to apply pre-processing on the images of the dataset (CLAHE filter), which turned out to have a slightly lower accuracy on the same architecture, with less 0,2%. With the best model, a simple application was created to allow the patients to use the machine learning model.

The machine learning model created in this thesis achieved an exceptionally high accuracy, compared with the literature. Also, it stands out form the rest because it uses a recent and lightweight architecture and has an application running on the frontend which allows everybody, even people without technical knowledge, to use the application.

Key words: melanoma; artificial intelligence; machine learning; image classification; image segmentation; supervised learning; medical imaging; MobileNet V2; Python.

Index

List of Figures

List of Tables

List of Abbreviations

SMM –

ALM –

NM –

ML – Machine Learning

AI – Artificial Intelligence

CHAPTER 1 – INTRODUCTION

1. Introduction

Melanoma has been increasing in white population in the past two decades [2]. It has been showed that the survival rate on melanoma heavily depends on how fast the diagnose is done as a consequence, early detection is critical [3]. However, there are some difficulties associated with recognition of melanoma, such as pigmented nevi (PN), seborrheic keratosis (SK), and other types of skin cancer such as basal cell carcinoma (BCC) [4]. Thus, the urge of new and faster methods of early detection are growing as the incidence rate grows. In the last couple of years many advances have been made in image processing and segmentation using several techniques of machine learning such as Multi Modal Neural networks or Convolutional Neural Networks and deep learning. Recent studies in digital skin diagnosis have used convolutional neural networks (CNNs) to classify images of melanoma and nevi with accuracies comparable with those achieved dermatologist-level diagnostic performance [3, 5].

To answer this demand, and in order to detect skin cancer in early stages, in this thesis, it is intended to create a Deep Learning algorithm that can predict skin cancer with precision and accuracy. The success of the Deep Learning algorithms depends heavily on the amount and quality of the dataset it learns. Hence, the first part of the thesis focuses on creating a complete dataset with a vast representation of melanoma, other various types of skin cancer and regular moles. With a complete dataset is possible to applicate Deep Learning techniques with the aim of creating a model that can predict from a simple image the probability of being a malign or benign, with precision and accuracy. The model will be running in the background of an application for an easier usage for the users.

With that in mind, this thesis is divided in four chapters, where in the first, it is presented the state-of -art of melanoma and Deep Learning applied to medicine and its developments and innovations in the last years. Following, it is presented the steps followed to create the dataset and the best model for this specific problematic. In the end, the results are discussed, and the measures used before are presented and justified.

CHAPTER 2 – LITERATURE REVIEW

2. Literature Review

2.1 Melanoma

As mentioned before, melanoma is a malignant neoplasm formed from melanocytes. The melanocytes are the unique cells with capacity to synthetize melanin within melanosomes. They can be found in many parts of the human body (suck as hair follicles, mucosa, cochlea, among other tissues) and in the case of the skin, these cells are exclusively derived from neural crest (temporary group of cells formed from the embryonic ectoderm germ layer, exclusively to vertebrates).

According to the World Health Organization there are many of subtypes known from melanoma. Many authors have abandoned the idea of melanoma classification as subtypes entirely since there are no alterations in the clinical management in the patient according to the subtype of melanoma [6] But, since there are some differences in the pathology of the various subtypes in this work melanoma subtypes will be considered. The most frequent and important are [7]:

i. Superficial Spreading Melanoma (SSM)
ii. Nodular Melanoma (NM)
iii. Lentigo Maligna Melanoma (LMM)
iv. Acral Lentiginous Melanoma (ALM)

SSM accounts for most cases in the ages between 30-50 years old, commonly on the trunk and on the legs on women and remains the most common subtype with about 70% of melanomas. NM is more observed in ages 40-60 years old and has a bigger incidence in men, almost two times higher. It represents about 15-30% of all melanomas. Lentigo and acral lentiginous has a lower incidence and represent less than 10% of all melanomas. 90% of the cases, melanoma develops on the face. LMM is particularly seen in cases of sunburned skin and has a higher incidence in 50-80 years old. 90% of the cases, melanoma develops on the face. ALM is seen on acral surfaces, reports about 75% of melanomas in non-Caucasian patients [8, 9, 10].

Melanoma classification and detection can be made with the unaided eye, for a more trained specialist. Nevertheless, unaided eye detection has been proven as not extremely precise, as seen in previous studies, with almost 80% precision among dermatologists and only 30% on non-dermatological specialists [11]. The use of health technologies as a skin microscope or a dermoscope allows improved visualizations of the lesions and consequently the improvement of the accuracy [12, 13].

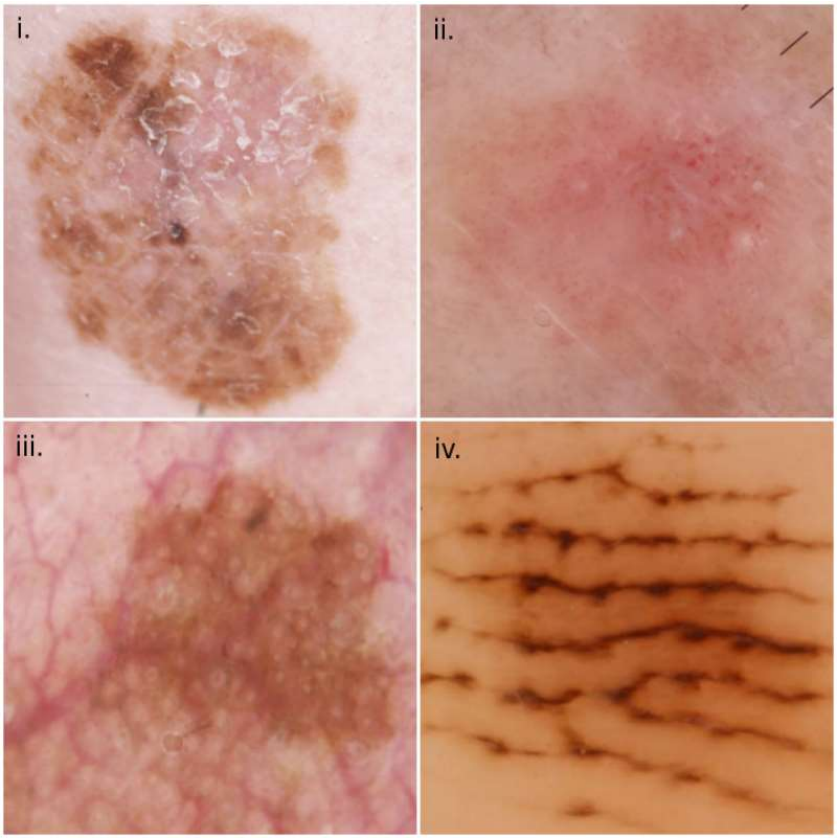


Figure 2.1 - Images from dermoscope of different subtypes of melanoma. i)Superficial Spreading Melanoma on Ankle. Source: Adapted from [14] ii) Nodular Melanoma form an approximately 75-year-old woman in a lower extremity taken from ISIC 2020 dataset. Source: Adapted from [15]. iii) Lentigo Maligna Melanoma, grey concentric circles seen from a New Zealand topic about dermatological problems. Source: Adapted from: [16] iv) Acral Lentiginous Melanoma, The parallel furrow pattern: in which parallel pigmented lines are seen along the furrows of the skin marks. Source: Adapted from [17].

Diagnosis of Melanoma can be made from a set of criteria called “ABCDE rule”, that tries to simplify the process of detection of melanoma. **A:** means asymmetry as usually melanoma is extremely asymmetric and irregular. **B:** means boarders and as said on asymmetry, melanoma tends to have irregular borders. **C:** for color, and in melanoma, opposite of “normal” skin moles, the color has more variations on melanoma. **D:** represents the diameter. It is a warning sign when the lesion size is about 6mm or larger.

E: stands for evolving or elevated surface. Nowadays, with growing information about melanoma and all skin related problems, the pursuance of the evolution is a big and important tool for the correct diagnose [11, 18].

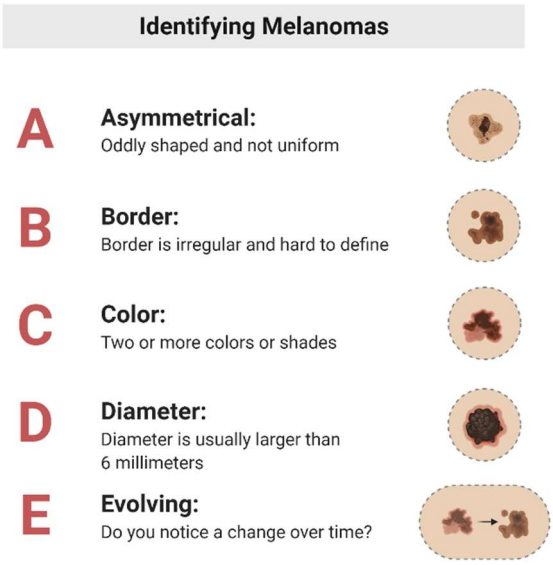


Figure 2.2 – ABCDE system for Melanoma diagnosis [19].

To diagnose the evolution of melanoma, a stage system was created that varies from stage 0 to stage IV. This is extremely important in mortality from melanoma, since the 5-year relative survival rate for patients with stage 0 melanoma is 97%, compared with about 10% for those with stage IV disease, thus the importance of early detection in this specific cancer.

The early classification of melanoma used to have based on the location on where melanoma appeared, but in 1960s Wallace Clark suggested a classification based on histological characteristics instead, entirely changing the way melanoma was diagnosed. Additionally, Clark proposed a system to evaluate the depth of invasion of melanoma cells on the skin. Clark, then, partitioned the skin into histologically identifiable anatomic compartments (Clark 'level' as they came to be known):

Level 1 – Melanoma cells are confined to the epidermis (melanoma *in situ*)

Level 2 – Invasion of a single cell or very small nests into the papillary dermis

Level 3 – Melanoma cells "fill and expand" the papillary dermis

Level 4 – Invasion into the reticular dermis

Level 5 – Invasion into the subcutaneous fat. [6, 19]

Clark, additionally, observed that patients with higher stages of melanoma were more prone to have lymph node invasion (Figure 2.2).

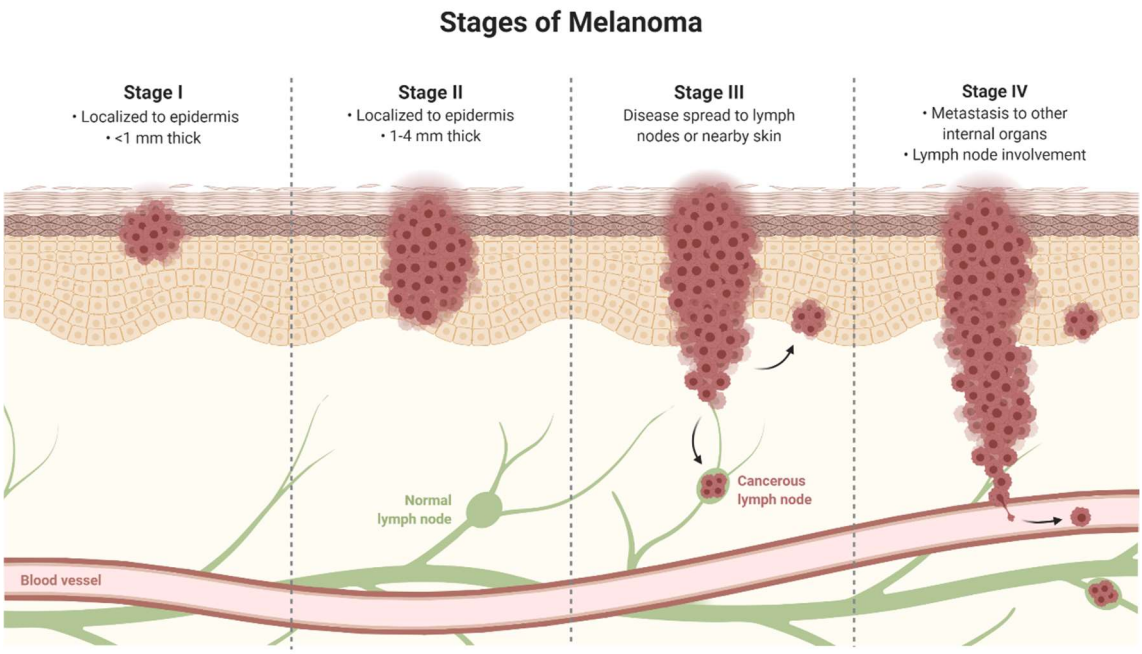


Figure 2.3 – Stages and progression of Melanoma [19].

However, in 1970, Alexander Breslow established a more precise method for diagnosing melanoma, based on a measured depth of invasion that captured the thickness of the tumor. Known as Breslow's depth, the classification system was based on the depth of invasion in millimeters rather than the depth by anatomic compartments and was initially divided into five stages [21]:

Stage I: less than or equal to 0.75 mm

Stage II: 0.76 − 1.5 mm

Stage III: 1.51–2.25 mm

Stage IV: 2.26–3.0 mm

Stage V: greater than 3.0 mm.

Breslow's system indicated that patients with early and thinner stages of melanoma had much higher probability of survival, and a lower chance of metastasis. Eventually, with the more advanced knowledge on melanoma, modifications of measurements had been made, and incorporated into staging systems used nowadays, represented on the figure above. Nevertheless, Breslow depth remains one of the best independent predictors of patient outcome [22].

Over the next years, as the data on patient treatment, survival rates and statistical methods improved, new staging systems were established. The tumor, node, metastasis (TNM) staging system was created by the American Joint Committee on Cancer (AJCC) for all major cancers, and melanoma is no exception [6, 21].

TMN staging system is being constantly reviewed according to new data and information about melanoma. The most recent TMN melanoma staging system edition was published and implemented in 2018 [23].

2.2 Epidemiology of Skin Cancer and Melanoma

Epidemiology identifies the distribution of the disease, as their factors underlining their source and causes. This gives a better understanding of the disease and therefore provides better information on how to control it. In the specific case of melanoma and non-melanoma skin cancer, the epidemiological study shows that the incidence differs substantially among different populations. The highest rates of incidence are demonstrated in Australia, North America and Northern and Western Europe. This significant variation is mainly attributed to a variety of risk factors such as ultraviolet (UV) exposure and phenotypic characteristics [24]. Largely due to the factors that Australia was colonized by fair skin people, and the high exposure to UV encountered in that region, at least 2 in 3 Australians will be diagnosed with some type of skin cancer before the age of 70. The equivalent is seen in the regions where melanoma tends to be the highest. Typically, fair skin populations in regions with exposure to UV radiation (Figure 4). On Europe, On Europe, the incidence of non-melanoma skin cancer and skin cancer grouped tends to be superior on western and northern Europe, where populations follow phenotypic characteristics described above.

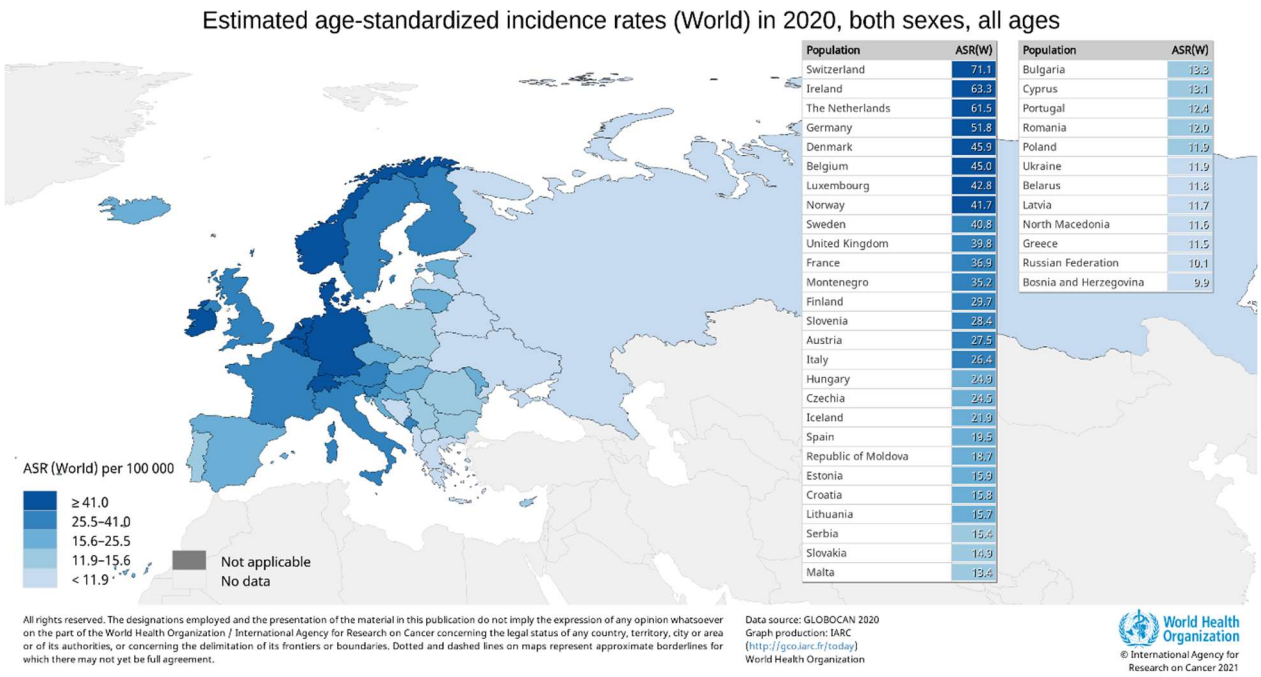


Figure 2.4 – Estimated incidence of Melanoma in Europe [25].

Malignant melanoma has a much more higher mortality rate than any other skin cancer, consequently the treatment of melanoma has a higher burden on society. The crude costs of malignant melanoma are expected to be around €2.7 billion for all EU regions in 2019 [26]. As the incidence in Europe are expected to grow 33.2% for 2040 the crude costs will too if there are no major advances either on treatment or prevention until then.

As for rest of the world, as expected, Oceania takes the first place in age-standardized incidence (ASI) with 174 cancer cases per 100 000 people. Europe takes a third place in incidence, but with higher mortality in both sexes then Northern America (Figure 2.5).

Estimated age-standardized incidence and mortality rates (World) in 2020, melanoma of skin, non-melanoma skin cancer, all ages

	Males		Females	
	Incidence	Mortality	Incidence	Mortality
Oceania	174.0	5.5	117.6	2.7
Northern America	108.8	2.5	57.6	1.0
Europe	34.7	2.8	22.8	1.6
World	18.9	1.5	10.9	0.88
Latin America and the Caribbean	12.6	2.0	8.7	1.1
Africa	4.8	1.7	3.9	1.4
Asia	2.2	0.85	1.6	0.53

Figure 2.5 – Estimated age-standardized incidence and mortality, in 2020, in all Continents, age 0-85+, by sex.

It is perceptible that ASI not only differs by regions, but also by sex, where the highest rates of melanoma are observed in males, almost in every region (Figure 2.6).

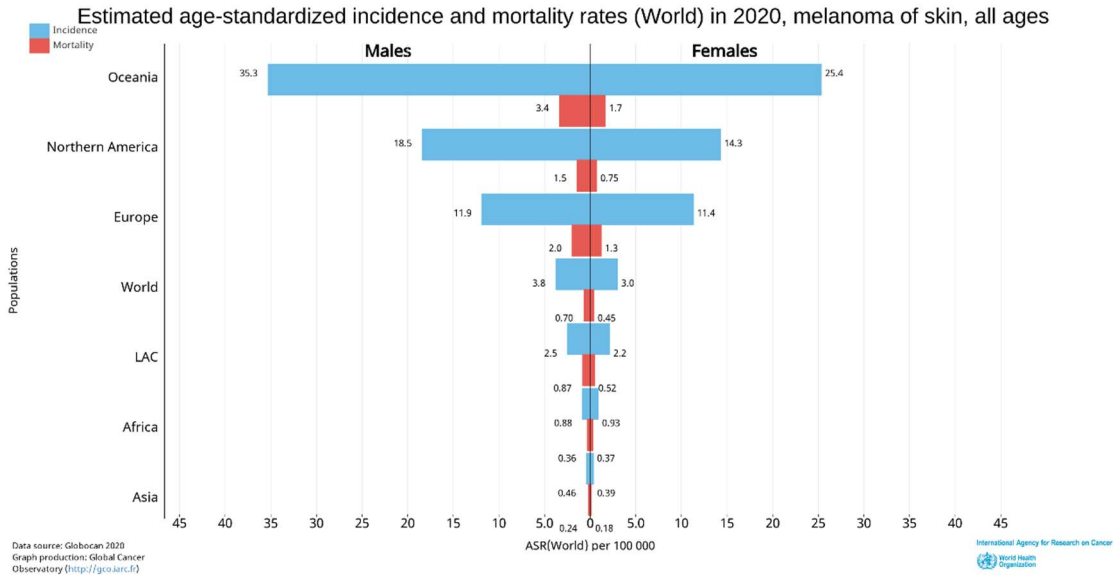


Figure 2.6 – Estimated incidence of Melanoma (worldwide) by gender [25].

However, in a study produced in 2020, the authors demonstrates that melanoma incidence is affected not only by sex but also by age [27]. The study cohort involved more than 5000 primary cutaneous melanomas and got to the conclusion that the median age for patients in melanoma is bigger for males. Also, to ages lower than 60 years, the incidence was even bigger for females. According to that study, the same thing can be seen around the world. The ASI when categorized by ages, < 45 years old and >= 45 years old, the ASI in the 1st group is significantly higher for females rather than males (Figure 2.7-a), the opposite happens for the 2nd group (Figure 2.7-b).

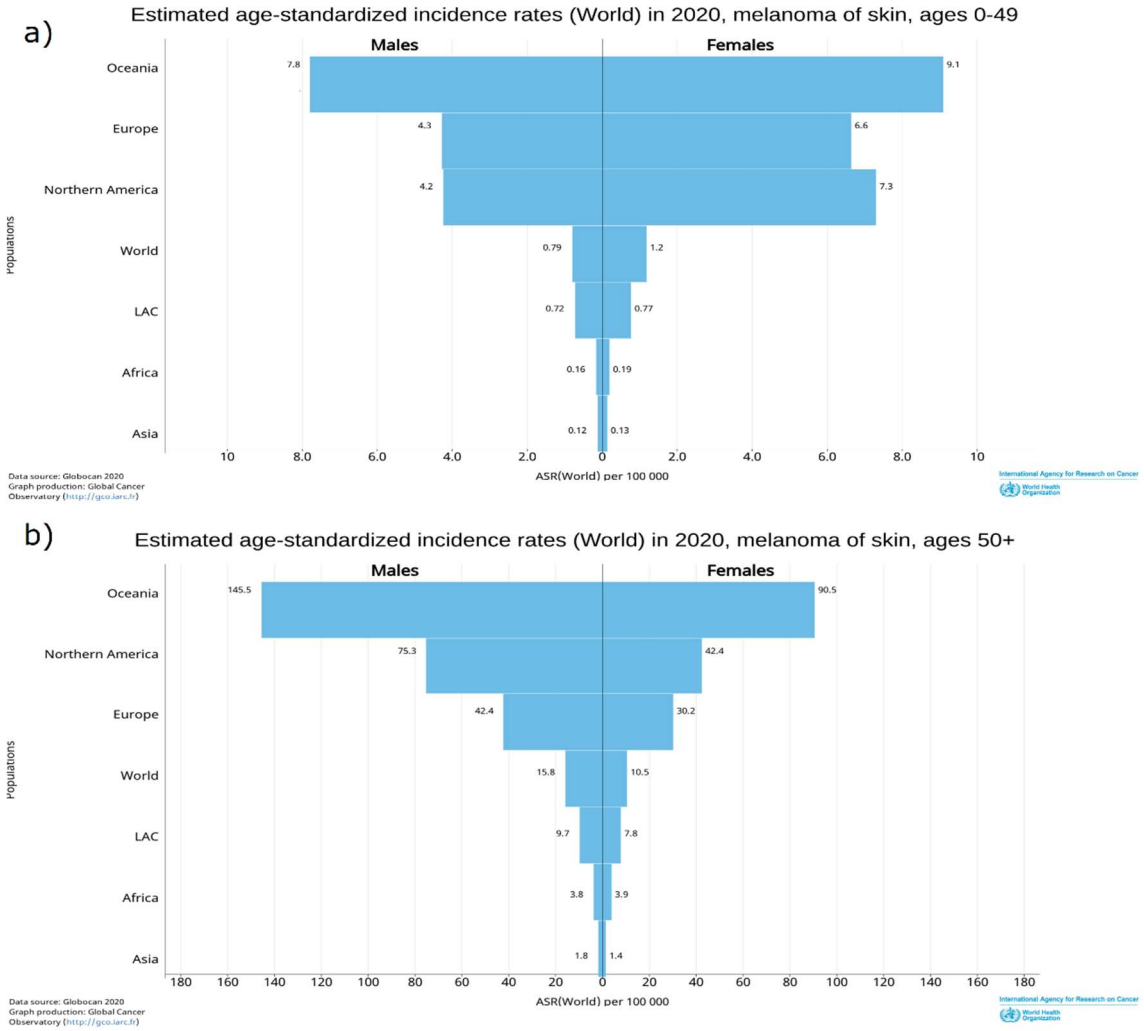


Figure 2.7 – Estimated incidence by world region and gender. A) age filtered from 0-49 and B) for 50+ years old [25].

Since, from the data shown, it is possible to understand that melanoma has been increasing since there are records until now. However, the projections for the future do not show any slowdown at all. On the contrary, the estimated number of deaths (both from melanoma and non-melanoma skin cancers) for 2040 show an increase of 5.71% in males and 5.4 % for females per year, compared to 2020. The estimated number of new cases continues the trend and has an estimated growth of 6.1 % for males and 5.8% for females (Figures 2.8 and 2.9). These projections are calculated assuming a growth of population of 1% a year. However, in 2020, the real growth rate is slightly higher than 1%, and the impacts of the ascending number of new cases and deaths can be higher.

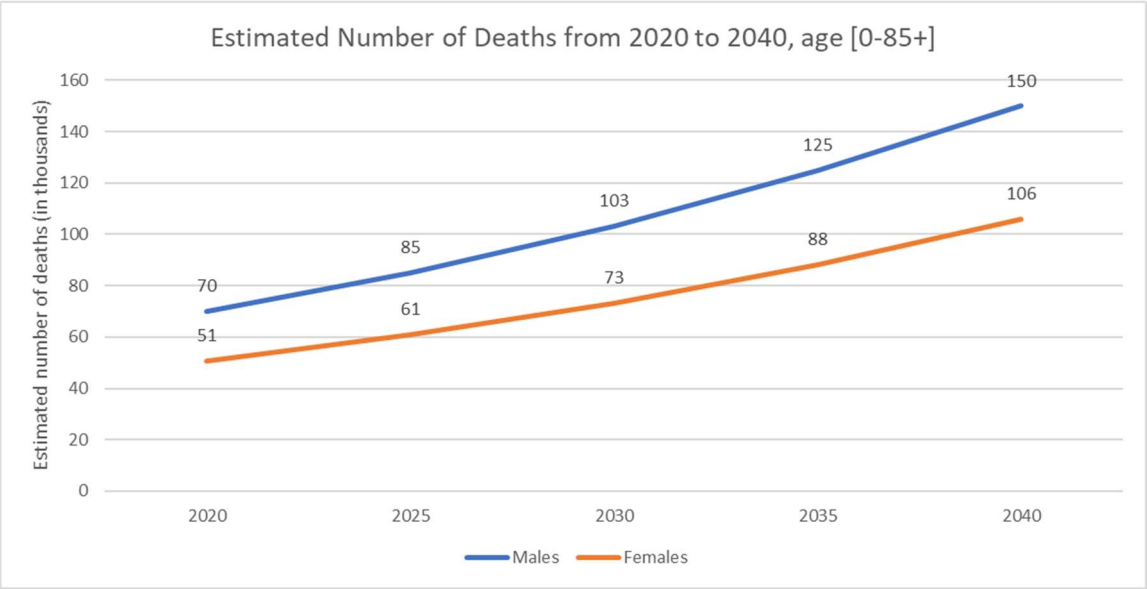


Figure 2.8 – Estimated number of deaths (worldwide) by gender from Melanoma and Non-Melanoma skin cancer [25].

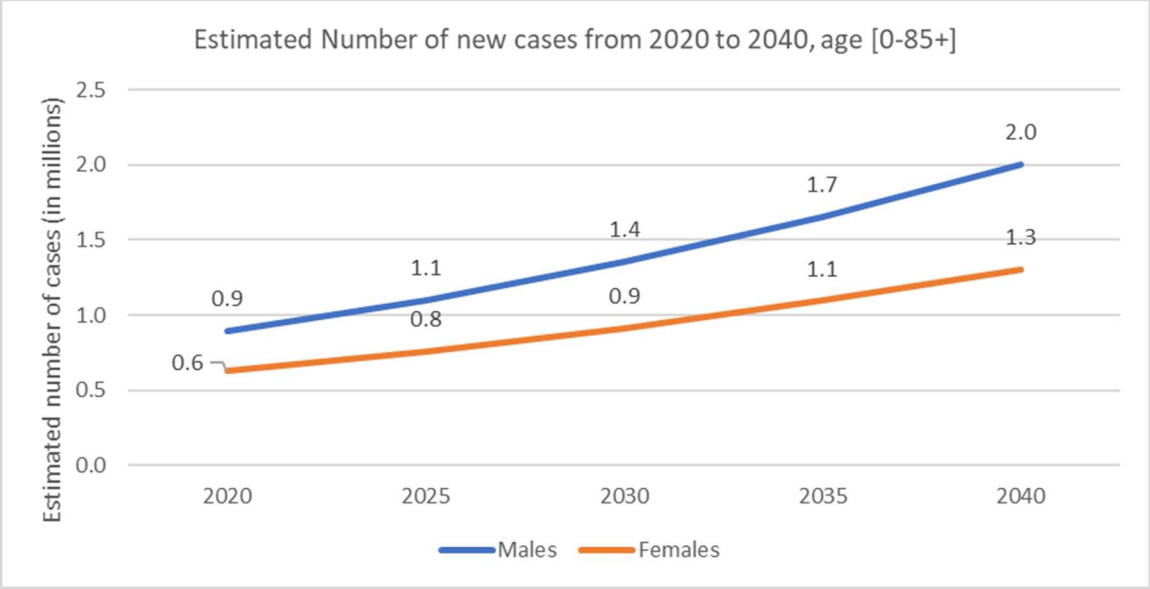


Figure 2.9 - Estimated number of new cases (worldwide) by gender from Melanoma and Non-Melanoma skin cancer [25].

2.3 Machine Learning

In recent years, deep learning has made breakthroughs in the fields of computer vision, speech recognition, natural language processing, audio recognition, medical imaging and bioinformatics [28].

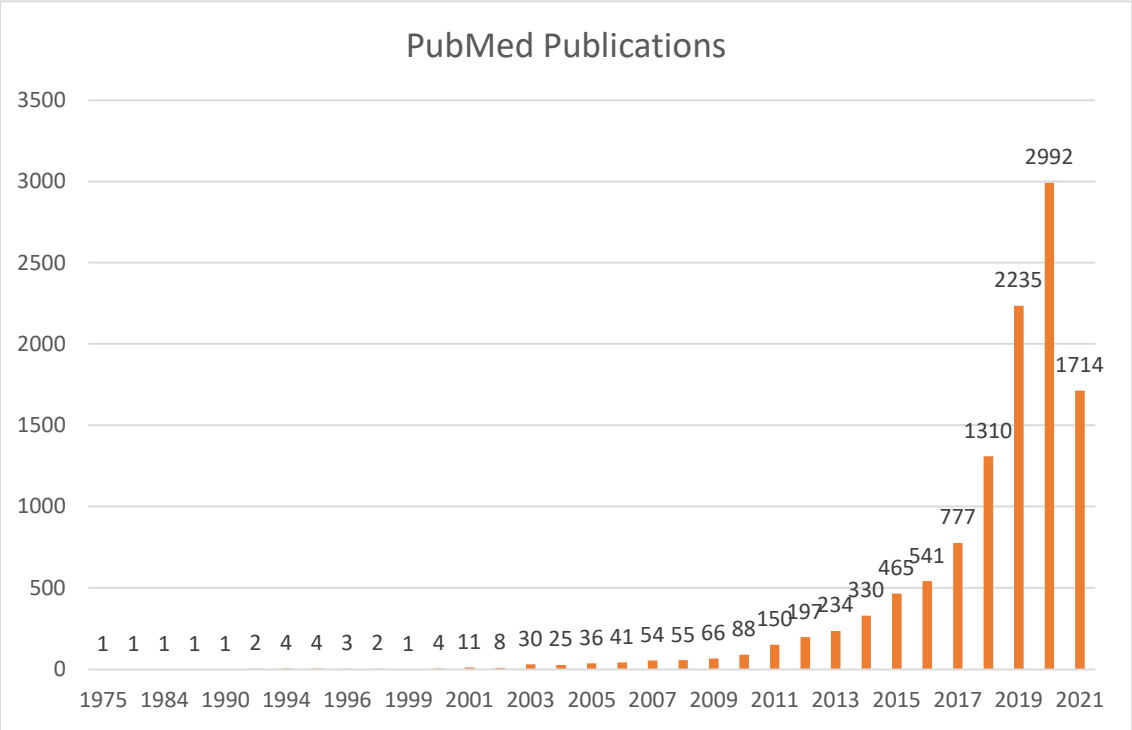


Figure 2.10 - Number of publications from 1975 to 2021 with the keyworks machine learning and medical diagnosis.

Medical care is about the health of people. At present, the amount of medical data is huge, but it is crucial to make good use of this huge medical data to contribute to the medical industry. Although the amount of medical data is huge, there are still many problems: medical data is diverse, including maps, texts, videos, magnets, etc.; due to different equipment used, the quality of data varies greatly; data presents fluctuating characteristics, over time and specific events change; due to differences in individuals, the law of the disease has no universal applicability [29]. This section first introduces the importance and evolution of machine learning and the application of deep learning algorithms in medical image analysis, expounds the techniques of deep learning classification and segmentation, and introduces the more classic and current ordinary neural network models. Following, a state-of-the-art on the application of AI for melanoma diagnosis and classification. Machine and deep learning algorithms play an important role to train the computer system as an expert which can be used further for prediction and decision making. Machine learning is the field of study that provides

computers the ability to learn without being explicitly programmed [30]. Deep learning is a type of machine learning that empowers systems to gain for a fact and comprehend the world regarding a pecking order of ideas. These fields bring intelligence into a computer that can extract the patterns according to the specific data and then process for automatic reasoning [28, 29]. Medical imaging is the rapidly growing research area that is used to diagnose a disease for early treatment. The function of image processing in the health domain is relative to the growing position of medical imaging.

The artificial intelligence is the main domain combined with machine learning and deep learning that work under this domain as shown in Figure 2.11. The AI is the major field to display human intelligence in a machine, machine learning is used to achieve artificial intelligence, while deep learning is a technique used to implement machine learning [33].

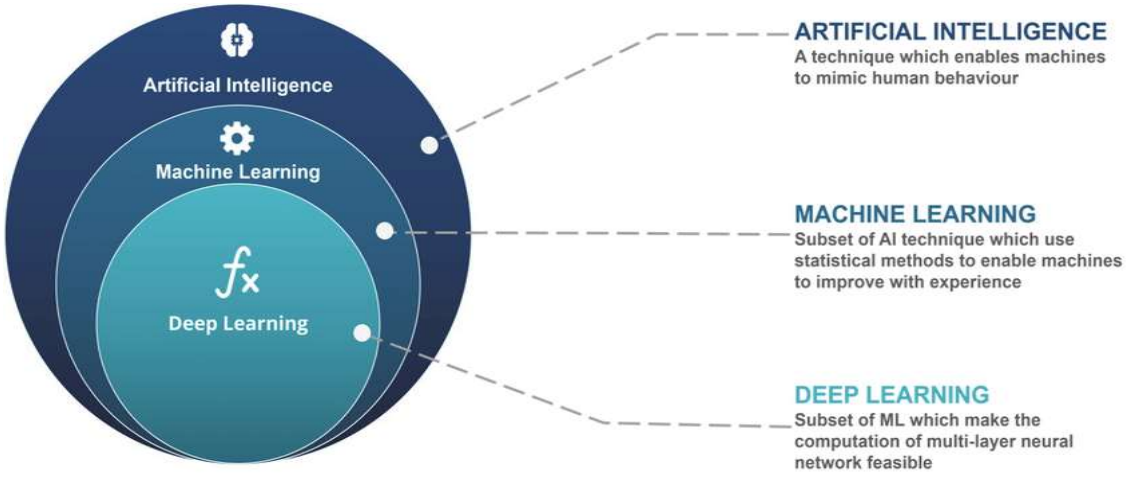


Figure 2.11 – AI, ML and DL domains [34].

When the creation of the algorithm is aimed on medical images, often, there are various steps performed on medical images before the detection of output. Initially, the medical image is inserted as input to the machine and deep learning algorithms. After that, the image is divided into different segments to zoom the interested area. Then, the features are extracted from these segments through information retrieval techniques.

Next, the desired features are selected, and the noise is removed (Data Preparation section on Figure 2.3).

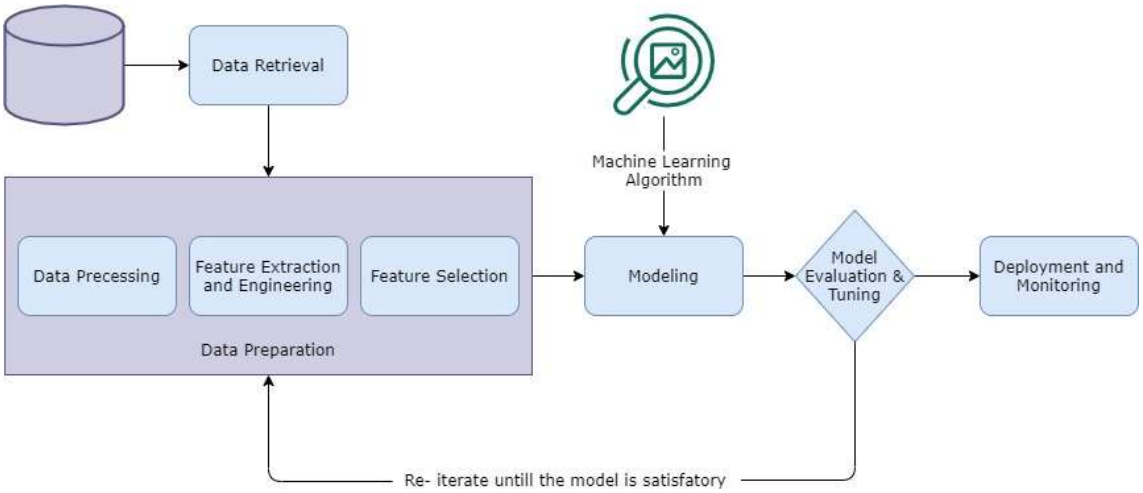


Figure 2.12 – Workflow diagram of Machine Learning algorithms in medical image.

Finally, the model is used to classify the extracted data and make predictions based on this classification. These steps are used in every experiment of machine learning. The supervised, semi-supervised, unsupervised, reinforcement and active learning algorithms are the main categories of machine learning. And last but not the least, the deep learning methods are essentially advanced phase of machine learning algorithms that classify data and predict more accurately using neural network [35].

2.3.1 Machine Learning Categories

2.3.1.1 Supervised Learning

It gives a training set of instances with appropriate objectives to a computer system. Taking this training se system give response accurately on given possible inputs. The classification and regression are the categories of Supervised Learning.

- The inputs are distributed into different classes using classification methods, and the trained system must generate actions that allocate hidden inputs to these classes. This is called multi labeling process. The spam purifying is the case of classification, in which the emails are classified into "spam" and, "not spam".

- The regression is a supervised technique in which the outcomes are continuous rather than discrete. The regression predictions are evaluated using root mean squared error (RMSE), unlike classification predictions in which accuracy is used as a performance measure.

Accuracy is the measure of all the correctly identified cases. It is most used when all the classes are equally important. In medical studies, it is often used since diagnosing a pathology has the same importance being positive or negative. The negative in medical models is extremely important since the patient is misdiagnosed and does not start treatment, which aggravates the pathology.

$$Accuracy = \frac{True\ Positive + True\ Negative}{(True\ Positive + False\ Positive + \ True\ Negative + False\ Negative}$$

2.3.1.2 Unsupervised Learning

The system will take the decision by itself rather train on the basis of some dataset. No labeling is given to the system that can be used for predictions. Unsupervised learning can be used to retrieve the hidden pattern with the help of feature learning of the given data.

- The clustering is an unsupervised learning approach that is used to divide the inputs into clusters. These clusters are not identified earlier. It builds groups on the basis of resemblance.

2.3.1.3 Semi-supervised Learning

In Semi-supervised learning, the system is assumed to be partial training data. This category is used with some trained data that can target some missing results. This kind of algorithm is used on untagged data for training commitment. The semi-supervised learning algorithm trained on both labeled and unlabeled data and this learning exhibits the features the features of both the unsupervised-learning and supervised learning algorithms.

2.3.1.4 Deep Learning

This is the advance phase of machine learning which mainly uses neural networks for learning and prediction of data. It is a group of different algorithms. These are used to design complex generalize system that can take any type of problems and give predictions. It uses the deep graph with numerous processing layer, made up of many linear and nonlinear conversion [36].

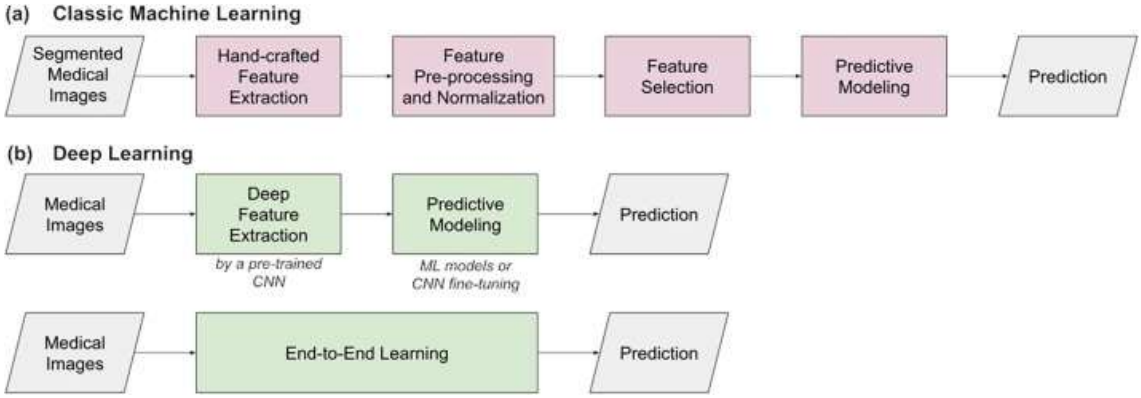


Figure 2.13 – Typical architecture and workflow of artificial intelligence systems for predictive modelling: a) classic machine learning, with the various processing steps involving hand-crafted features such as in radiomics; b) deep learning considering either deep medical image feature extraction or end-to-end learning [37].

2.3.2 Deep Learning Architectures

Deep learning has developed into a hot research field, and there are dozens of algorithms, each with its own advantages and disadvantages. These algorithms are widely used and often focus on image classification and segmentation tasks.

2.3.2.1 Convolutional Neural Network (CNN)

CNN is a deep learning algorithm, which means a reliable algorithm for recognition of large amounts of data and processed with high-speed machines. CNN layers are composed of the input layer, hidden layer, and output layer. Since Krizhevsky et al. proposed AlexNet based on deep learning model CNN in 2012 [38], which won the championship in the ImageNet image classification of that year, deep learning began to grow. In 2014, GoogLeNet and VGGNet both improved the accuracy on the ImageNet dataset [39]. GoogLeNet has further developed the v2, v3 and v4 versions to improve performance [40]. In 2012, the AlexNet adopted an 8-layer network structure consisting of five convolutional layers and three fully connected layers. After each convolution in five convolutional layers, a maximum pooling performed to reduce the amount of data.

AlexNet accepts 227×227 pixels' input data. After five rounds of convolution and pooling operations, the 6×6×256 feature matrix finally sent to the fully connected layer. This model's error rate in ImageNet dataset was 15.3%, which was much lower than the second placed model, with 26.2%. The major difference is that its activation function is not sigmoid but adopted ReLU, and, given the results, proved that the ReLU function is more effective.

VGG16 and VGG19 were first proposed by Visual Geometry Group (VGG) of Oxford University in 2014, and since then, it has been one of the most successful architectures for medical imaging. VGG16 network architecture has 41 layers and VGG19 has 47 layers. Compared with AlexNet, VGG uses smaller kernels.

For a given receptive field range, the effect of using several small convolution kernels is better than using a larger convolution kernel, because the multi-layer nonlinear layer can increase the network depth to ensure more complex patterns are learned, and the computational cost is smaller. VGG simplifies the process by creating 3 × 3 filters for each layer. The use of uniform and smaller filter sizes on VGG can produce more complex features and lower computing when compared to AlexNet [41].

Numerous convolutions and initiations are followed in CNN which has pooling layers and utilizes backpropagation relating to conventional neural networks.

1) Convolutional layers: Makes a component guide to anticipate the category probabilities for every element by applying a channel that filters the whole picture, scarcely any pixels all directly.
2) Activation layer: Activation layer downsizes the measure of knowledge the convolutional layer produced for every element and keeps up the foremost basic data.
3) Pooling: component map is created based on the information obtained from convolutional layer and commonly pooled.
4) Fully connected input layer: straightens" the outputs produced by past layers to rework them into a solitary vector which will be utilized as an input for the subsequent layer.
5) Fully connected output layer: Produces the last probabilities to make a decision a category for the image [42].

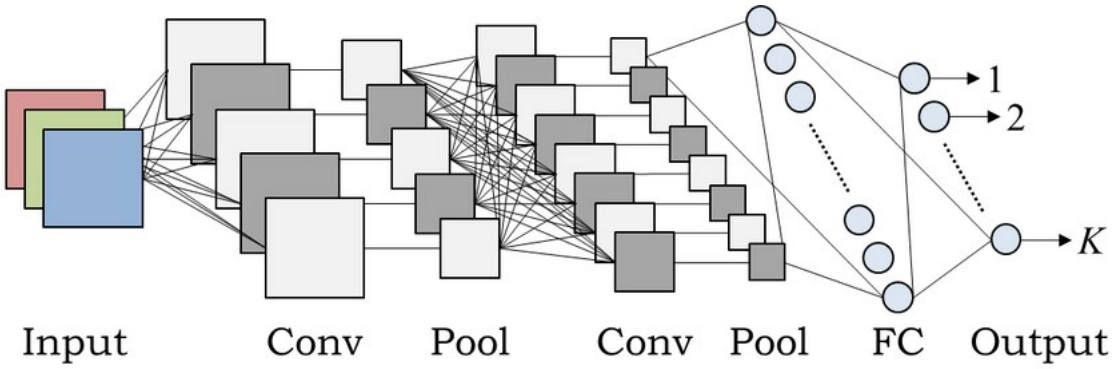


Figure 2.14– An example of CNN architecture [43].

2.3.3 Deep Learning in Medical Imaging

With the development of deep learning, computer vision uses a lot of deep learning to deal with various image problems. Medical image as a special visual image has attracted the attention of many researchers due to its importance and large-scale availability. To guide computers to learn features that can characterize the data for the given issue, the idea lies at the foundation of several deep learning procedures. The most used type of deep learning models for image analysis is convolutional neural networks (CNNs) [44].

2.3.3.1 Classification

The classification if medical image is the main task in deep learning in order to investigate for clinical-related issues for early treatment of the patient. Anthony J. et al., showed a fine-tuned model that evidently outdid feature extraction, attaining 57.6% accuracy in multiclass score evaluation of knee osteoarthritis against 53.4% [45]. However, the Kim et al. shown that by CNN feature retrieval performed fine-tuning in cytopathology image classification accuracy with 70.5% versus 69.1% [46].

2.3.3.2 Segmentation

Semantic segmentation is an important research field of deep learning. The segmentation process is used to process the organs and substructures of the medical images. It is used for quantitative analysis of the clinical features, such as brain and cardiac analysis. Since CNN's success in the classification field, people started to try CNN for image segmentation and with the development of deep learning technology, some segmentation networks based on convolution structure are derived. The fully convolutional network (FCN) [47] proposed by Long et al. is one example.

It replaces the fully connected layers of the classification network VGG16 with convolutional layers and retains the spatial information of the feature map and achieves pixel-level classification. Finally, FCN uses the deconvolution and fusing feature maps to restore the image and provides the segmentation result of each pixel by softmax. The U-Net, proposed by Olaf [48] is the combination of upsampling and downsampling layers architectures. It merged the connections of convolution and de-convolution samples of layers [49]. The performance of the FCN on the Pascal VOC 2012 datasets [50] has increased by 20% compared to the previous method, reaching 62.2%.

Alternatively, deep transfer learning techniques, which relax the hypothesis that training and testing data comes from the same probability distribution, allow avoiding training DL models from scratch. Deep transfer learning techniques have been classified into four categories: instances-, mapping-, network-, and adversarial-based, as detailed by Tan et al. [51]. This technique is widely used in CNNs like VGG16 and VGG19.

2.3.4 Deep Learning Development Frameworks

While the deep learning technology is developing in theory, the software development framework based on deep learning theory is also booming.

2.3.4.1 TensorFlow

TensorFlow is an open-source software library that uses data flow diagrams for numerical calculations. Google officially opened the computing framework TensorFlow on November 9, 2015, and officially released Google TensorFlow version 1.0 in 2017. The TensorFlow calculation framework can well support various algorithms for deep learning such as CNN, RNN and LSTM, but its application is not limited to deep learning, but also supports the construction of general machine learning. It can run on multiple GPUs at the same time and can automatically run the model on different platforms. Because TensorFlow developed in C++, it has high-performance.

2.3.4.2 PyTorch

PyTorch is specifically targeted at GPU-accelerated deep neural network programming. PyTorch's calculation graph is dynamic, and the calculation graph can be changed in real-time according to the calculation needs. PyTorch immediately attracted widespread attention as soon as it was launched, and quickly became popular in research

2.3.4 Deep Learning in Melanoma Classification

In the last years, machine learning (ML) applied to medical diagnosis has been growing rapidly since this methods allow to characterize and utilize large and complex data sources for diagnosis and predictions [52]. Traditional methods of programming have a prerequisite of well-defined patterns to mathematical approach to solve the initial problem. An algorithm is a well-defined set of rules that is used to solve the initial problem with a finite number of steps. The big step back on this approach is that biology, frequently has various patterns therefore is extremely hard to create a set of rules to diagnose medical problems, including melanoma and skin cancer in this study. Even if melanoma respects perfectly the ABCDE rule (Figure 1.2), is almost impossible to create a set of rules for an algorithm to diagnose it. Therefore, ML is a very attractive solution for biological programming problems.

There are already several approaches to implement an automated method for melanoma detection, and each author has different perspectives on how to address the problems that are universal to the melanoma (and other cancers) detection and diagnosis. The first step to implement a machine learning model for classification involves the acquisition of the tissue digital image. The most common techniques used for this purpose are the epiluminence microscopy (ELM), it is also known as dermoscopy [53]. Although these are the most common, with the advance of smartphone cameras over the years, smartphone datasets of melanoma can encourage experts to use them for automated diagnosis [54].

In some important related work, it is possible to conclude that different methods to diagnose melanoma can lead to decent results on more than one method. Nikhil Cheerla et al. [55] , suggested an algorithm emphasizing the texture and local pattern for melanoma segmentation. After their application this classification method returns a sensitivity of 97%. Using a different technique, Margarida Ruela et al. [56], addressed the classification for melanoma and non-melanoma images.

For their approach they implemented Image segmentation, via manually (through the help of an expert) and automatically, by using adaptive thresholding algorithms. Before the lesion classification, in this work, it was implemented some feature extractions as simple shape and symmetry-related features, moment invariants and Fourier descriptions. The final accuracy obtained on their model was 96.8%, for the PH^2 dataset [56]. Some authors tend to combine the two main ways of developing a computer-aided diagnosis (CAD), and choose to go for a hybrid model. Jinen Daghrir et al. for example, used an hybrid model in their work that consists on a convolutional neural network and two classical machine learning classifiers trained with a set of features describing the borders, texture and the color of a skin lesion [57]. In this work the dataset used was the ISIC ((International Skin Imaging Collaboration), an archive that contains more than 23000 images all of them token with the dermoscopy technique. On more recent years, as the neural networks became more and more advanced, the feature extraction and pre-processing of the dataset becomes not as much relevant as before. In a recent work published in 2020, using the same dataset (ISIC) the authors Qishen Ha et al assemble a convolutional neural network (CNN) model that without any feature extraction was able to achieve around 96% of accuracy [58].

2.3.5 Aim of the Study

With this study, it is intended to create a mobile application that uses AI (artificial intelligence), more specifically ML (machine learning) to diagnose melanoma out of a photograph taken by the user's smartphone. For that it is essential to understand what advances had been made with ML and image processing and diagnosis. Other stage of major importance is creating a complete and large dataset. Since in this chapter all melanoma developments and characteristics were aborded, in the next chapter a state of the art on machine learning and its medical applications will be introduced. Following that, the methods and results achieved in this project will be presented and discussed.

CHAPTER 3 – MATERIALS AND METHODS

3. Materials and Methods

3.1 Methodology

There are 3 steps that define a machine learning development workflow, in terms of methods:

1. Data Acquisition
2. Data pre-processing
3. Researching the best model for the type of data and model construction

The basic workflow follows these 3 steps, where in a first glance, it is necessary to compile all data into one good and strong dataset, where the different models are trained and tested, in order to select the optimum parameters and choose the best model for this specific task. After this step, when the final ML model is reached, the application is built to provide the implementation and the access of the ML algorithm to the user. In that way, the user is just accessing to a website, uploading an image, and getting a quick response from the ML model. The general workflow for this thesis is shown in the image below (Figure 2.1).

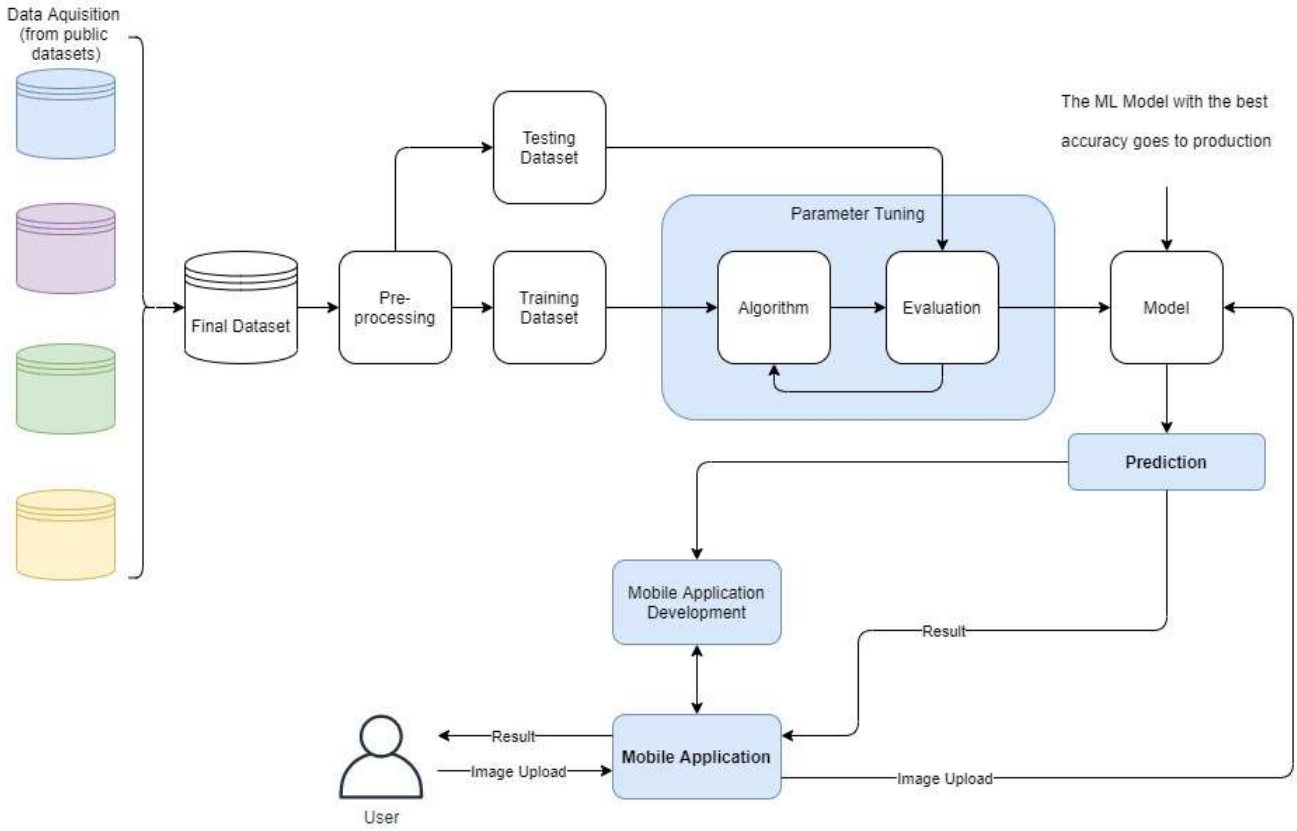


Figure 3.1 – Workflow diagram of machine learning algorithm and mobile application development.

3.2 Data Acquisition and ML Model construction

3.2.1 Data Acquisition

One of most important factors to create a good machine learning model is the quality of the data used. Since the ML model is expected to predict skin cancer, it is essential to collect as many images of skin moles, both benign and malignant, as possible. Not only the images are essential, but the diagnosis of the mole in the image, is extremely relevant to have a precise predict, since the ML model will connect the image with the diagnosis given and learn about the features in that image. Therefore, it was created a set of rules to choose datasets available for this study. The dataset needed to be free and available online, must have the diagnosis (by biopsy) of the mole for every image in the dataset, all the datasets should have the same format of images (JPG/JPEG, preferably), and finally each dataset needs to have a metadata file where is possible to get the diagnosis made on each image. The metadata file is important since to create a ML model it is necessary to organize the data into what it is called as labels (which is what the model will assign to the images in that directory). If there is no metadata it is not possible to validate that each image is on the right directory and that can create massive errors in the model.

Nowadays, the amount of data available online is enormous, and a few datasets were able to satisfy the pre-requisites. The datasets used in the ML Model are listed in the Table 3.1.

Table 3.1 – Datasets used to create the ML model and important characteristics.

Dataset Name	Number of Images	Image Size (px)	Image Format
ISIC	> 44000	1024x1024	jpeg
HAM 10000	10000	600x450	jpg
PH2	199	765x562	bmp
Smartphone Dataset	1342	1050x1050	jpg

Since the datasets that fulfill the initial prompt were selected, it was needed to start to organize the dataset that is going to be used in this study, it is indispensable to understand the dataset chosen and comprehend what it is possible to extract from each. With the help of several Python libraries (such as Pandas, NumPy, Matplotlib, Seaborn,

etc.) it is possible to apply data science techniques to visualize the data of these datasets with a more thoughtful analysis.

3.2.1.1 ISIC archive

ISIC (International Skin Imaging Collaboration) is an industry partnership designed to facilitate the application of digital skin imaging to help reduce melanoma mortality. The goal of this archive includes a large and expanding open-source public access archive of skin images. [15] All the images on this dataset are taken with the dermoscopy technique. There are more than 150,000 images in this archive but only almost 70,000 were made public. In this work, it was used about 40,000 images. For this dataset, it was downloaded through Kaggle, which is a subsidiary from Google®, and offers a community of data science and machine learning with ready to use datasets [59].

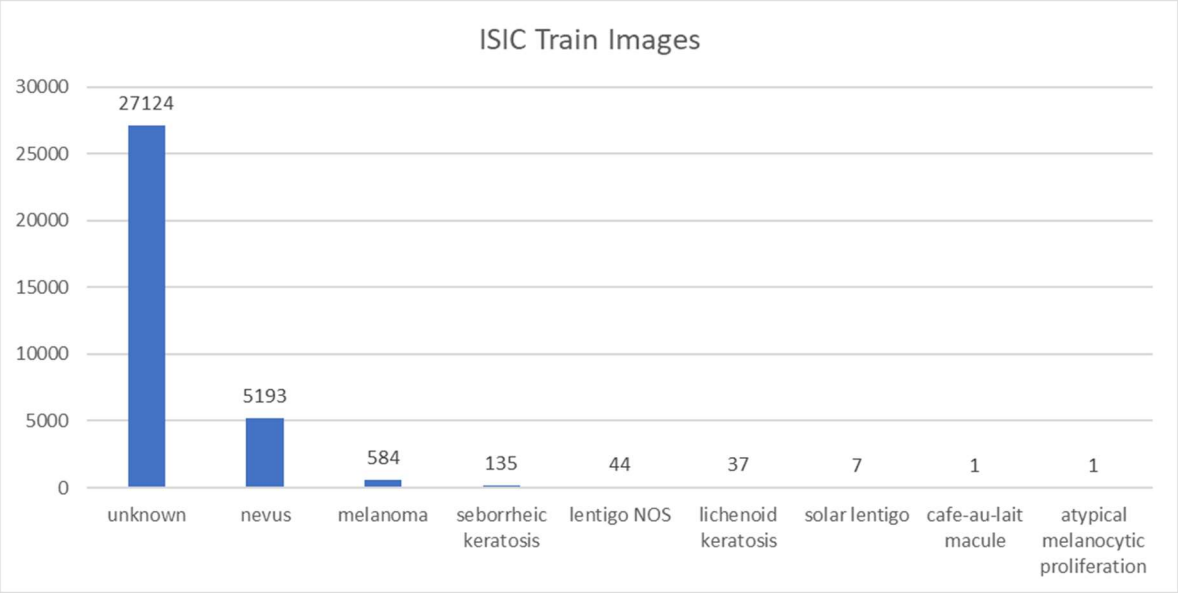


Figure 3.2 – Number of images per category taken from the metadata file of the ISIC dataset used.

ISIC in Kaggle® has the images divided by 2 directories, Train and Test, as this is conventional in ML since the model needs different paths for images to make sure the model is not tested in the same images as it trains. In this dataset, the Test folder has no metadata, and therefore was not used in this study. Also, one important note about this data, is that all the "unknown" images are benign skin moles, but were in fact verified by biopsy as well as the other diagnostics.

3.2.1.2 HAM10000

The second biggest dataset used in this study is the HAM10000 (Human Against Machine with 10000 training images) dataset [60]. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include with the proportions as seen in the figure 3.2:

- Actinic keratoses and intraepithelial carcinoma / Bowen's disease “akiec”
- Basal cell carcinoma - “bcc”
- Benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses) – “bkl”
- Dermatofibroma - “df”
- Melanoma - “mel”
- Melanocytic nevi “nv”
- Vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage) – “vasc”

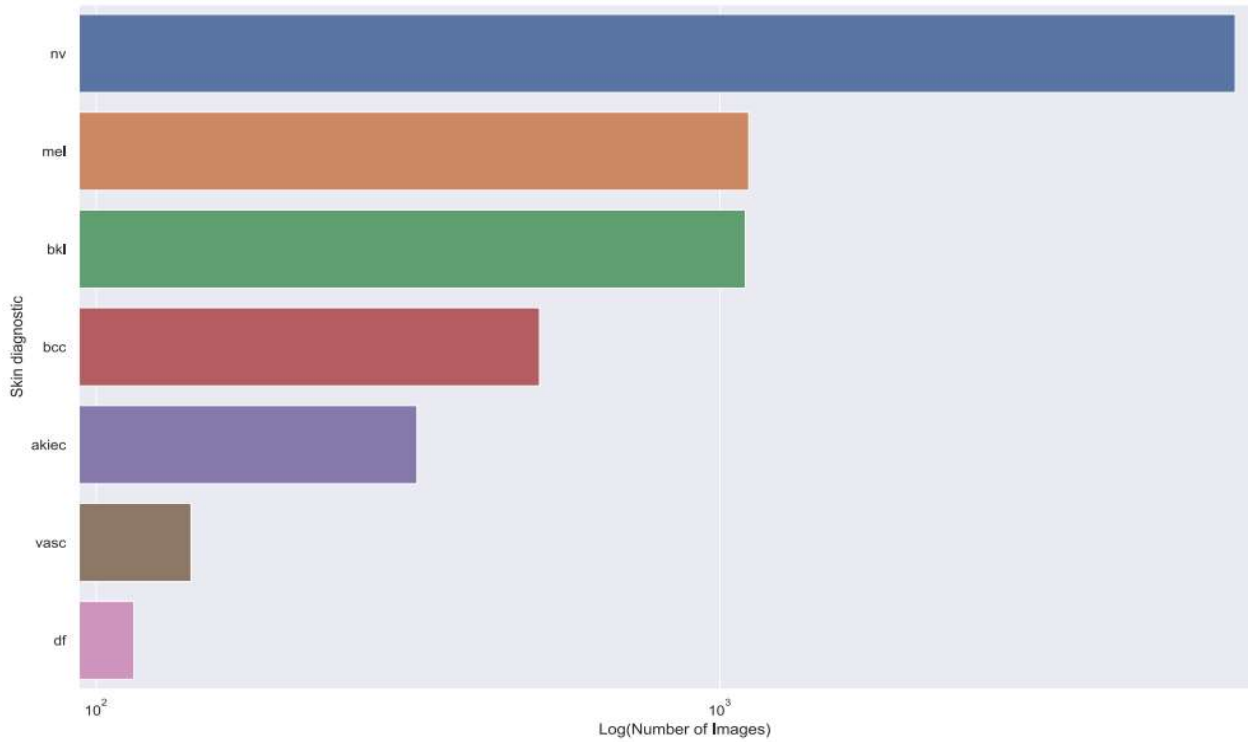


Figure 3.3 – Proportions of the logarithm of the number of images by skin diagnostic from the HAM10000 dataset. This plot was taken in one of the data science tasks made on this project with Python ®.

3.2.1.3 Smartphone dataset

All the images in previous datasets were taken using a dermatoscope. Although, the aim of this study, as said before, is to create a mobile app that can output a predict from an image of a skin mole. However, the common user does not have a dermatoscope laying around, and the most probable object to use for this case is the smartphone. Since the images taken from a smartphone are not close to those taken with dermatoscopy. Therefore, a dataset with smartphone images of skin moles was extremely important in this study since it provided images closer to the reality. It was found a dataset that fulfill all the initial requirements for this study, and it was published by Thanh-Toan Do et. al. [61]

In this dataset, the authors used biopsy to validate the outcome of the skin moles. However, not all the images were in this condition. But since this type of images were important to this study, the images of the skin moles that were biopsied were selected. The dataset had a total of 2298 images and to start to understand the data it was splitted into categories. The first and most important was if the skin mole of the image was in fact confirmed with a biopsy, as seen in Table 3.2, being the "true" category the selected for this study.

Table 3.2 – Number of images separated by biopsy. True is the ones that in the study a biopsy was realized.

Biopsy	Images
TRUE	1342
FALSE	956

Since all the images that belong to the category of false were removed the number of images used from this dataset was 1342. Subsequently, to continue to comprehend the data it was then divided by diagnostic. Again, with the help of Python® it was plotted the number of images by diagnostic (Figure 3.3). Since there are more than one thousand pictures and the diagnostic follows the pattern of the datasets chosen before, it was selected for this study.

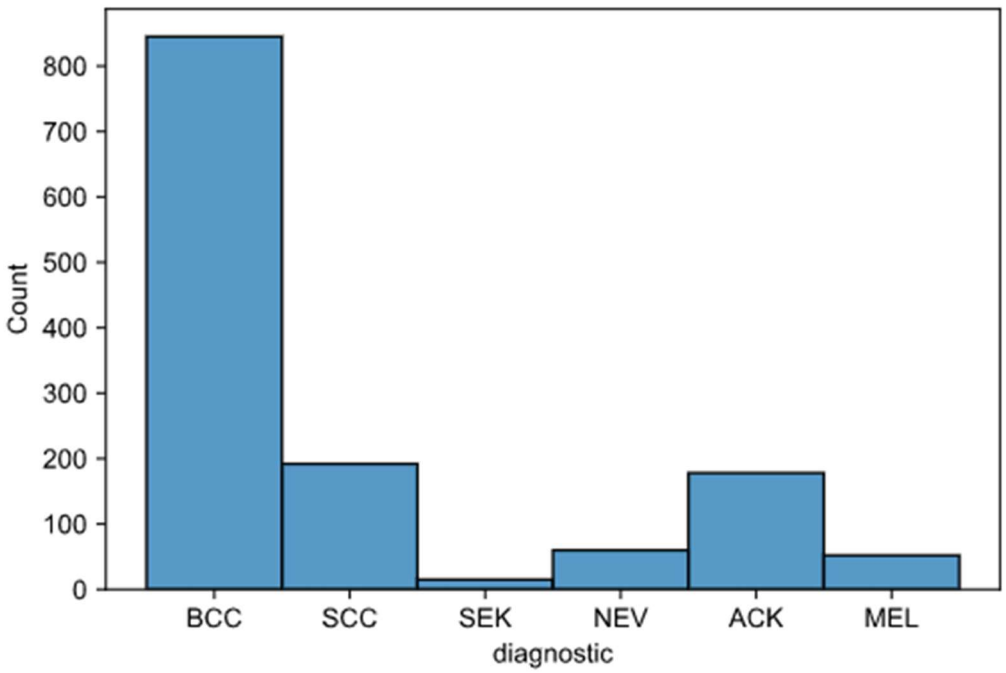


Figure 3.4 – Plot of the count of the images in the smartphone dataset. The type of diagnostic includes BCC - Basal cell carcinoma; SCC- Squamous Cell Carcinoma; SEK - Seborrheic Keratosis; NEV – Nevus; ACK - Actinic keratosis and MEL – Melanoma. This plot was taken in one of the data science tasks made on this project with Python ®.

3.2.1.4 PH² Dataset

At this point of the study, the number of images selected was more than enough to create the ML model. Therefore, it was necessary to have a dataset to validate the data in a mass scale. For that purpose, this dataset fits perfectly. The PH² dataset has been developed for research and benchmarking purposes, facilitate comparative studies on both segmentation and classification algorithms of dermoscopic images. PH² is a dermoscopic image dataset acquired at the Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal [62]. With almost 200 images it has 80 images of common nevus, 80 of atypical nevus and 40 of melanoma. PH² dataset has many more features for the images besides diagnostic, such as asymmetry, color, pigment network, and many more. Since the ML model in this study will not evaluate those features, it was just collected the images and the diagnostic for each image.

3.2.2 Final Dataset used in this study

3.2.2.1 Data Compilation

With all the datasets chosen, it was necessary to concatenate all the data with the images and features chosen from the datasets. Using data science techniques, since the amount of data is enormous a set of rules was created to ensure all the data was divided into the right paths with the right diagnostic. This step was made with this step was made with small advances, and tested at the end of each one, to ensure data reliability. Since all the metadata files were different from each other it was also necessary to create some ground rules so when all the info is aggregated there is no mistakes. One extremely important rule was to ensure that there are no duplicates in the same dataset and on others, to not have repeated images. Since both HAM10000 and ISIC are from a competition online, it was discovered that both have some repeated images. Those images were eliminated and only 1 copy was used.

3.3 Pre-Processing

3.3.1 Train/Test split

All the images were segmented into 2 categories, benign or malignant and by diagnostic. after the dataset was concatenated, it was randomly split into 2 folders. 80% of the images were for model training and 20% for testing. Since model training would be performed on online servers, this step is only done once on a local computer and does not have to be repeated every time a new model is tested. In addition, the same images are always used to train the various ML models and allows for a direct comparison between them, that would not possible if random separation was done for each model. with this method of random separation an associated problem arose. What happened is that as the separation was random, some of the diagnoses were much more representative either in the training folder or in the test folder. This is not great, because in some diagnoses where there are fewer images, the training folder, which serves as the basis for the ML model to train, would have almost no data. Thus, the separation of data was done randomly (with the same data science techniques in Python) but ensuring that each diagnosis has the same proportion of data both in training and in testing. Therefore, the proportions of the dataset created in this study for train and test can be seen in the Figure 3.5.

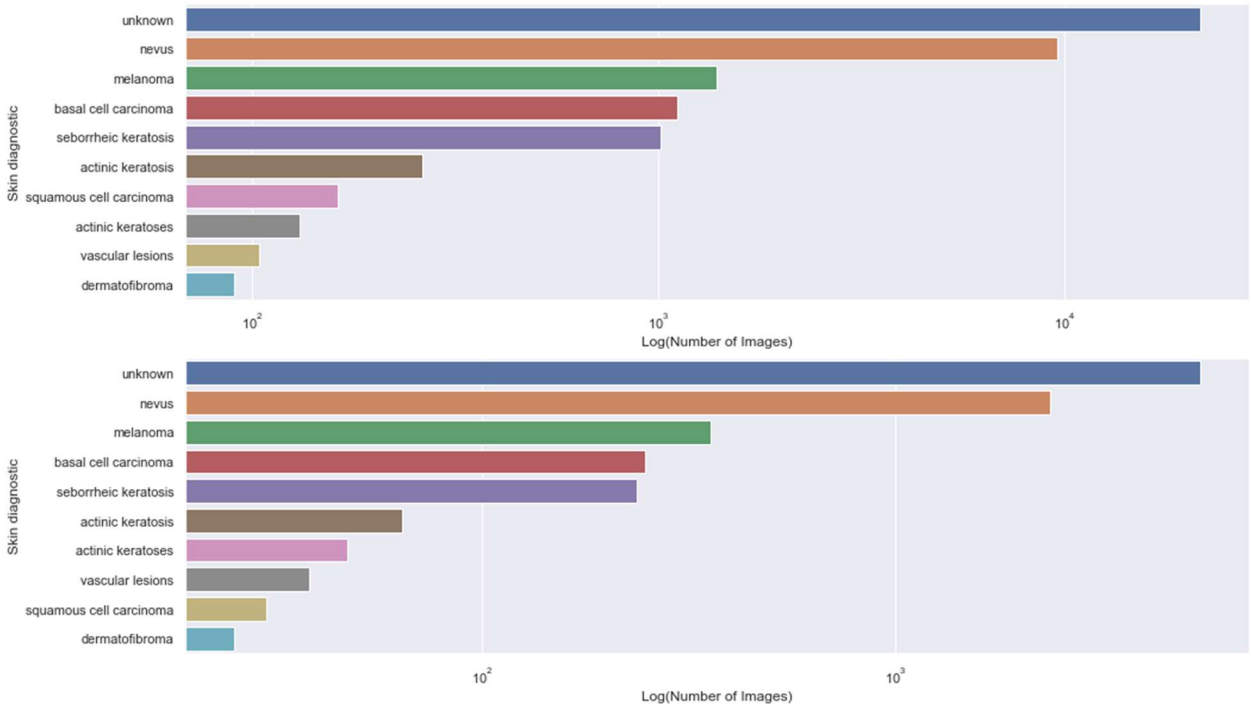


Figure 3.5 – Plot of the outcome of the separation of the images by the folders "Train" and "Test" where it is noticeable that the proportions of the images are the same. Above we verified the proportions of the images in logarithmic scale of the "training" images divided by the diagnosis. Below we check the same for the "test" data. Although the proportions are the same, the number of images is not (Train - 80% of total), therefore the logarithmic scale on both plots is different. This plot was taken in one of the data science tasks made on this project with Python ®.

3.3.2 Unbalanced Data

Nevertheless, the dataset was not yet ready to use. If we look at the proportions of the several diagnostics in the datasets used in this study, it is perceptible that there is no data balance. Since the performance of ML models significantly decreases with unbalanced data, this is an essential point for the study [63]. It can be seen in Figure 3.6 that the discrepancy between malignant and benign images is extremely pronounced.

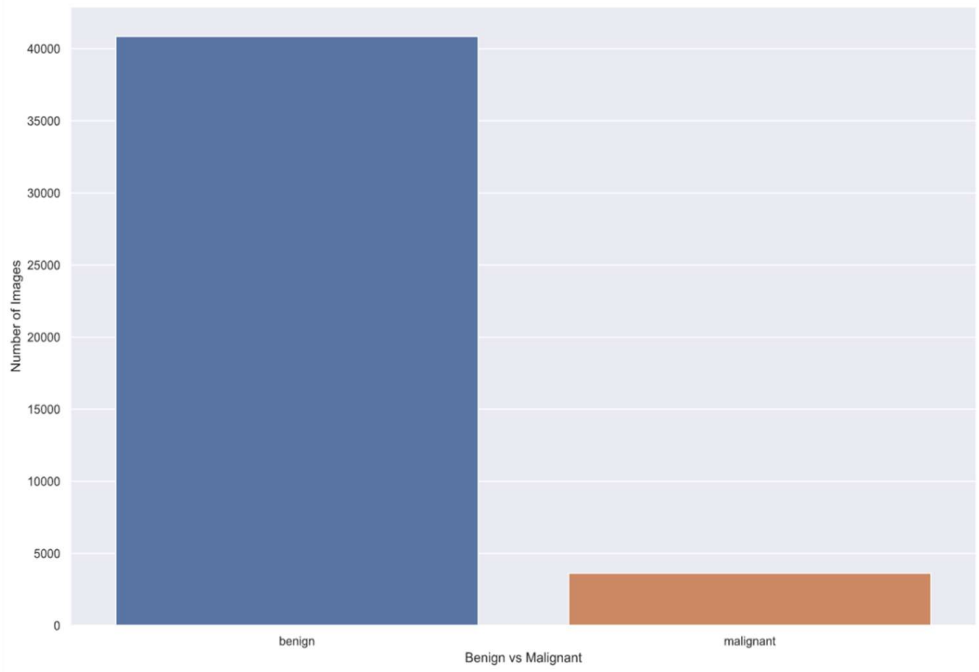


Figure 3.6 – Bar chart of the number of images on the dataset used in this study, segmented by benign or malignant. This plot was taken in one of the data science tasks made on this project with Python ®.

To solve this problem there are 2 options, or we remove data from the class with the largest (benign) representation to equal the smallest, which is a faster and more efficient method, but you lose a substantial amount of data from the dataset. As an important part of this study was the collection and processing of data on a large scale, this option is not ideal at all in this study. The other option, which is more complicated and requires more advanced data processing, is to generate new data from what we have already obtained. This technique is known as data augmentation for image classification.

That is, if in the class that has less representation new data is generated from the existing ones, it is possible to obtain the same balance of dataset data and have the various classes with the same amount of data. This second option is better suited for this study. Since the ratio of the difference between the images of the 2 categories is approximately 10, if in each malignant image 9 new images are generated, the amount of data will be similar. This data augmentation is made from translations and rotations of the original image. In this study, a function was also introduced that allowed to vary the brightness of the image, to further simulate images taken with smartphones in favor of dermatoscope images where a perfect brightness is verified (Figure 3.7). This step allowed the dataset where the ML model is created to be balanced and the results optimized.

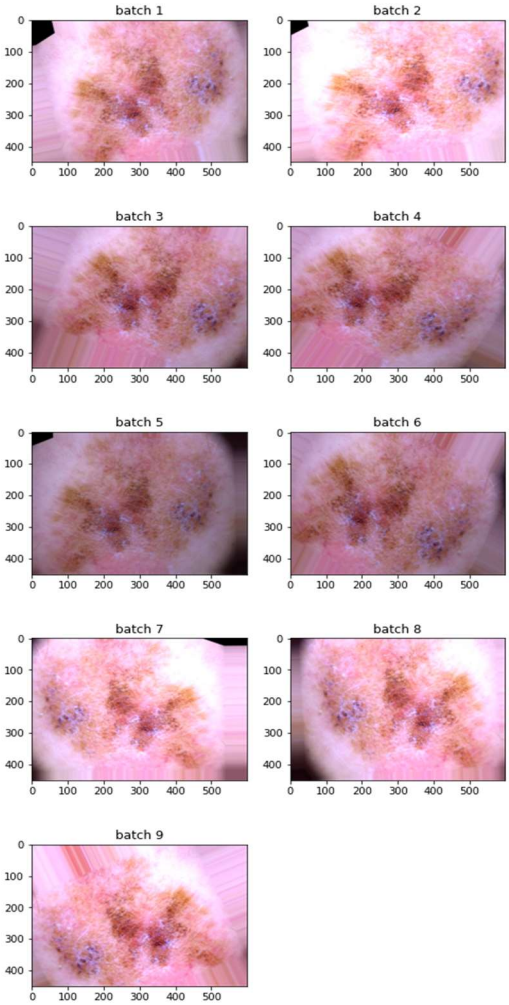


Figure 3.7 – Example of data augmentation for one image of the dataset. Each batch is randomly generated from a set of rules, created for this study. This plot was taken in one of the data science tasks made on this project with Python ®.

The various versions of the dataset were uploaded to the Kaggle® website (mentioned above) which provides space and resources so that models can be built, trained and tested with tools vastly superior to the usual computers.

3.3.3 Enhancement filter

In an attempt to test the quality of the images, a filter was applied into the dataset. The idea behind this step, is once the best model is chosen to go forward, it is again tested with this dataset in which the images have a pre-processing, and thus comparing the results with the original dataset. If there are no major changes then pre-processing is not necessary to solve this problem, and time is better invested in fine-tuning the chosen model.

To improve the contrast of the images, a copy of the complete dataset was created with an image enhancement filter applied. Several filters were tested, such as:

- Adaptive Thresholding
- Histogram equalization
- CLAHE (Contrast Limited Adaptive Histogram Equalization)

Since the time it takes to copy the complete dataset is high, the method of choosing the best of these filters was made with excerpts of small images (taken from the dataset) and compared with the naked eye. The filter that showed the best results was the CLAHE filter. (Figure 3.8). This algorithm works by creating several histograms of the image and uses all of these histograms to redistribute the lightness of the image. CLAHE can be applied to greyscale as well as color images.

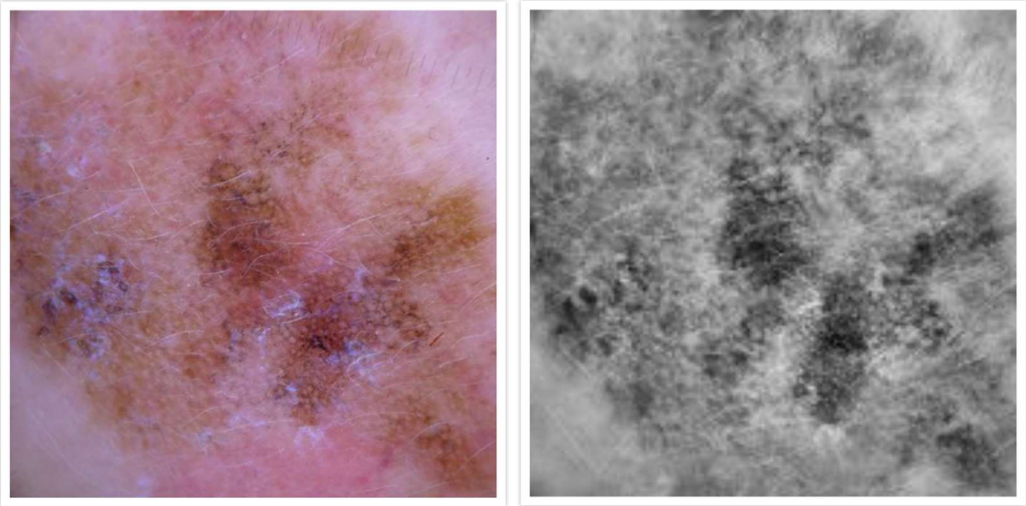


Figure 3.8 – Example of the CLAHE filter applied to an image from the original dataset, created for this study. The original image on the left and CLAHE filtered on right. The difference in colors is much more pronounced. This plot was taken in one of the data science tasks made on this project with Python ®.

3.4 ML Model Construction

For this project 3 different ML models architectures were chosen. VGG16 and VGG19 are architectures commonly associated with image processing and classification of various medical-related problems. Usually with these 2 architectures, the results are quite positive when the models are fine-tuned. Hameed et al., in their study, obtained predictions of 97.73% for VGG16 and 95.29% for VGG19, in the detection of breast cancer through histopathological images. [64] The third model used in this project is MobileNetV2. In some studies, such as the study by Roslidar Roslidar et al., an ML model of MobilenetV2 with fine-tuning can achieve an accuracy close to 100% in breast cancer as well [65].

3.4.1 VGG16

The VGG architecture has been widely used in computer vision over the last few years. It consists of stacked convolutional and max pooling layers. The VGG network was introduced by the researchers at Visual Graphics Group at Oxford (thus the name VGG). VGG Network has some pros and cons [66]:

Pros:

- It is a very good architecture for applying on a particular task, like image classification.
- Also, pre-trained networks for VGG are available freely on the internet, like the VGG-16 used in this project (Figure 3.9).

Cons:

- The main drawback is that it is very slow to train if trained from scratch. Even on a decent GPU, it would take days get it to work.

In this project, the ML models were trained and tested on Kaggle servers. Additionally, as the dataset is quite complete and contains a large number of images, the development of a CNN with this architecture was time-consuming. Although using a pre-trained VGG-16 architecture and with the help of high-performance GPUs provided on the servers (NVIDIA TESLA P100 GPUs). The development of each VGG-16 model took a few days, with integrated model development, training, testing and validation.

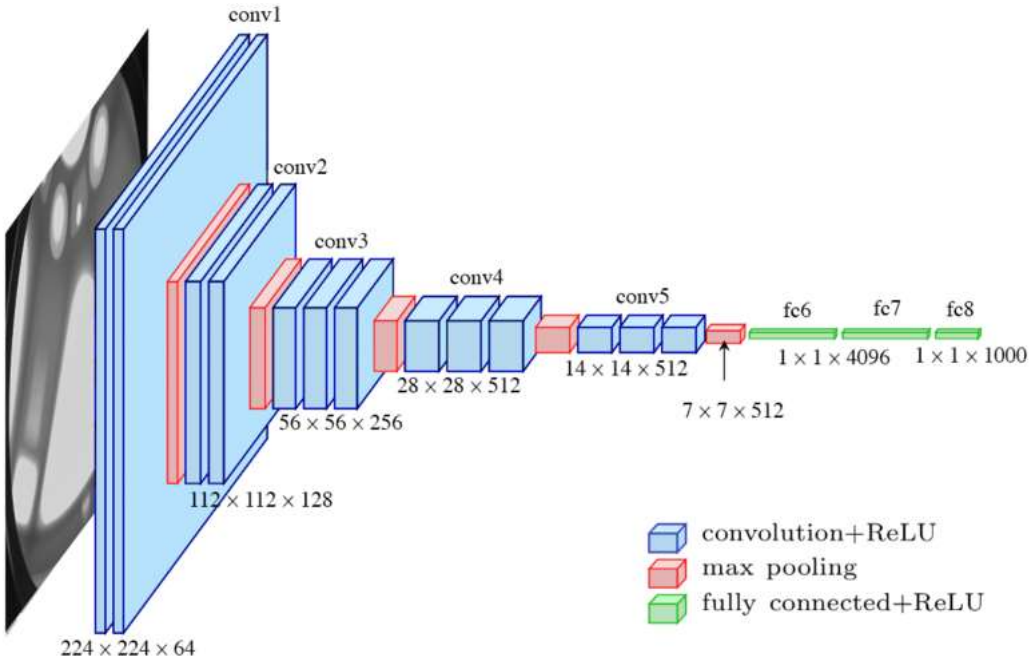


Figure 3.9 – Visual representation of the architecture of a pre-trained VGG-16, as used in this project. Convolution, mac pooling, fully connected and ReLU are types of layers. Image taken from Qassim et al. [67].

In this image, it is possible to visualize the 16 layers that compose the model (blue and green layers). Pooling layers provide an approach to down sampling feature maps by summarizing the presence of features in patches of the feature map. Two common pooling methods are average pooling and max pooling that summarize the average presence of a feature and the most activated presence of a feature respectively [68]. The rectified linear activation function (ReLU) is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero.

3.4.2 VGG19

Compared to VGG16, VGG19 is slightly better but requests more memory, since it has more layers (Figure 3.10). As with the VGG16, this model also takes a few days to be fully developed.

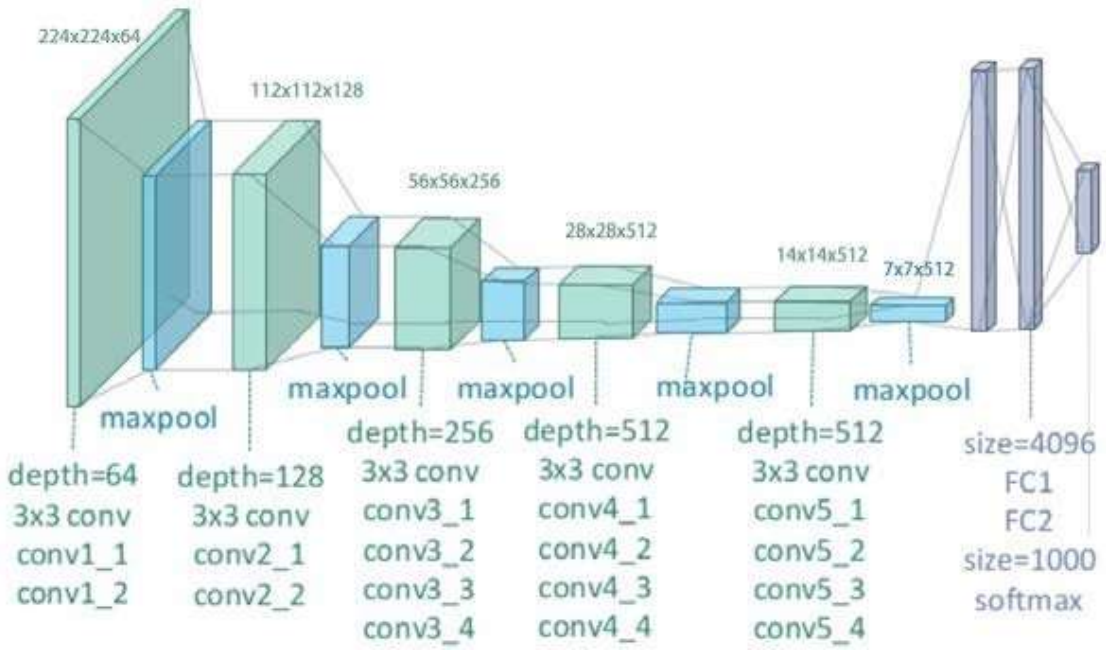


Figure 3.10 – Illustration of the network architecture of VGG-19 model: conv means convolution, FC means fully connected. Adapted from Clifford K. Yang et al. [69].

3.4.3 MobileNetV2

MobileNetV2, created by Google®, offers big functionality, like the previous models, but with much lower requirements. In this architecture, Depthwise Separable Convolution is introduced which dramatically reduce the complexity cost and model size of the network, which is suitable to Mobile devices, or any devices with low computational power. For this project on a first approach seems more suitable, since the amount of data is massive, and the development times can reduce. In MobileNetV2, there are 2 types of blocks that form the ML model (Figure 3.11). Compared to VGG models, mobileNetv2 is lighter and faster to train, not degrading functionality and accuracy.

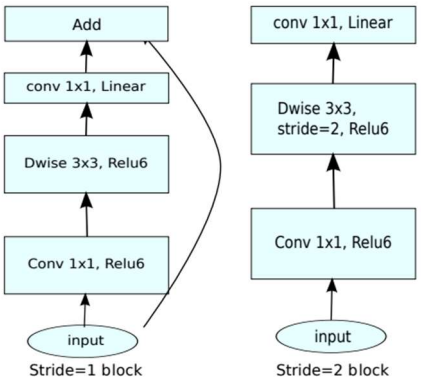


Figure 3.11 – Illustration of the MobileNetV2 architecture. Image taken from M. Sandler et al. [70].

3.4.4 Comparison between the chosen models

Table 3.3 – Direct comparison between the 3 chosen models and their respective implementations.

	MobileNetV2	VGG16	VGG19
Input Shape	(224, 224, 3)	(224, 224, 3)	(224, 224, 3)
Weights	ImageNet	ImageNet	ImageNet
Optimizer	Adam	RMSProp	Adam
Loss function	`Binary Crossentropy (from logits)`	Binary Crossentropy	Categorical Crossentropy

3.4.5 Production Model Architecture

With the study of the dataset done before, it was realized that for this project an ML model of Multi-Label Classification would have to be implemented, since there are more than 2 classes in the dataset. However, in this study the ultimate goal is not to identify the diagnosis of all skin moles. The aim is to detect and classify skin cancer and melanoma. Looking at the literature, binary classification problems generally have better results [71]. Thus, meeting the objective of the study, it is possible to implement 2 ML models using binary classification instead of a single multi-label classification model (Figure 3.12). Thus, it was chosen that, in a first phase, a model classifies the soft skin as benign or malignant, thus being a binary classification. If the template predicts malignancy, then that image is sorted by another template (more specific in detecting melanoma) to classify the skin mole as melanoma or non-melanoma skin cancer.

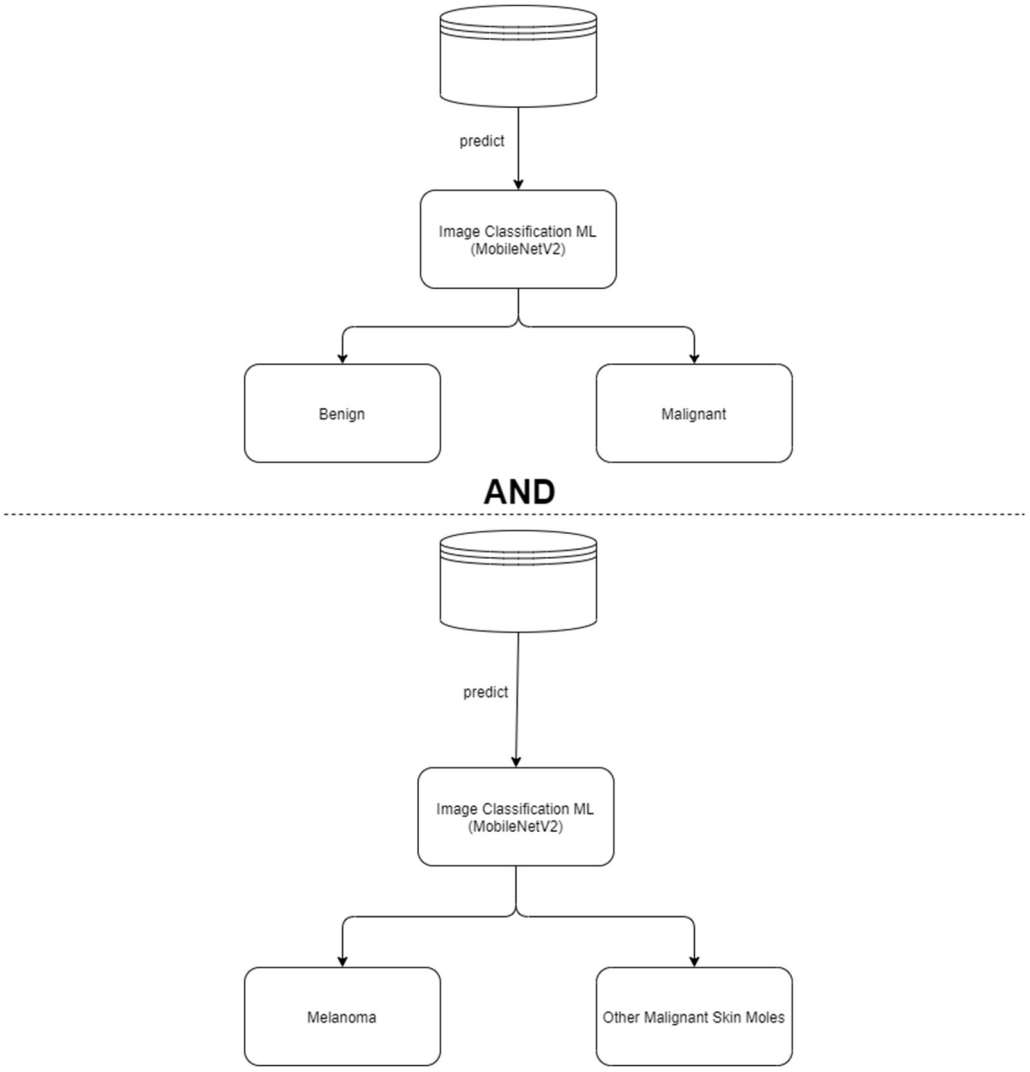


Figure 3.12 – Representative flowchart of the structure of the neural networks implemented in this project.

3.5 Mobile Application Development

The purpose of creating an application in this project is so that the user can effortlessly have access to the latest advances in machine learning applied to medicine. Since the use of ML models and implementation is not common for most users, an application with a simple interface that can use ML in the background is essential. In this project, since the development of ML models was done only with Python, it was decided to use a framework in the same language that would allow creating an application available on several platforms. There are a few possible ways to build the application, it is possible to use a combination of programming languages, each with its function, and assemble everything into one single stream of information (website e.g.). However, there is also some frameworks that can facilitate the construction in this case. In short, a framework is a set of ready-made elements, rules, and components that simplify the process and increase the development speed. Because of its code reusability and efficiency, it allows a programmer to bootstrap an application in no time and start implementing the business logic. Software frameworks come in different types for different programming languages. For Python, the programming language chosen in this Thesis, there are 2 frameworks that stand out from the rest, Flask and Django.

3.5.1 Flask

Flask is a micro-framework designed to create a web application in a short time. It only implements the core functionality giving developers the flexibility to add the feature as required during the implementation. It is a lightweight framework. This framework can either be used for pure backend as well as frontend if need be [72].

3.5.2 Django

Django is a framework for web development using python and created for the quick development of database driven sites [73]. It is a high-level web framework which allows performing rapid development. The primary goal of this web framework is to create complex database-driven websites.

3.5.3 Django and Flask comparison

Django:

- It is fully featured with database interface, application directory structure and administration panel
- The framework is in the market with a bigger active community
- It saves time, as it provides a built-in template engine
- Does not require third-party tools or libraries
- Increased security
- Apt for large projects

Flask:

- It is lightweight, flexible and simple
- Carries a customizable structure
- Great for small projects
- Easier learning curve
- Offers room for experimentation.

The framework chosen was Flask. The reason for choosing Flask over Django, is because Django is a very feature-heavy framework, ideal for large Web projects, while Flask is a light microframework. Given the nature of the project, Flask was chosen over Django for its simplicity and to avoid any unnecessary framework overhead.

3.5.4 Application Design

In this project, the design and beauty of the application were far less important than the usability and functionality. The basic idea is that the user can log into their account (or create one if it's the first time) and from there, upload one or more images to the neural network and get the result. In the figure below it is possible to see the application mockup in a simple way in a smartphone view. One of the advantages of having chosen a framework to create the application is that we only have to create it once, as the framework recognizes the device the user is accessing and adapts the application to it.

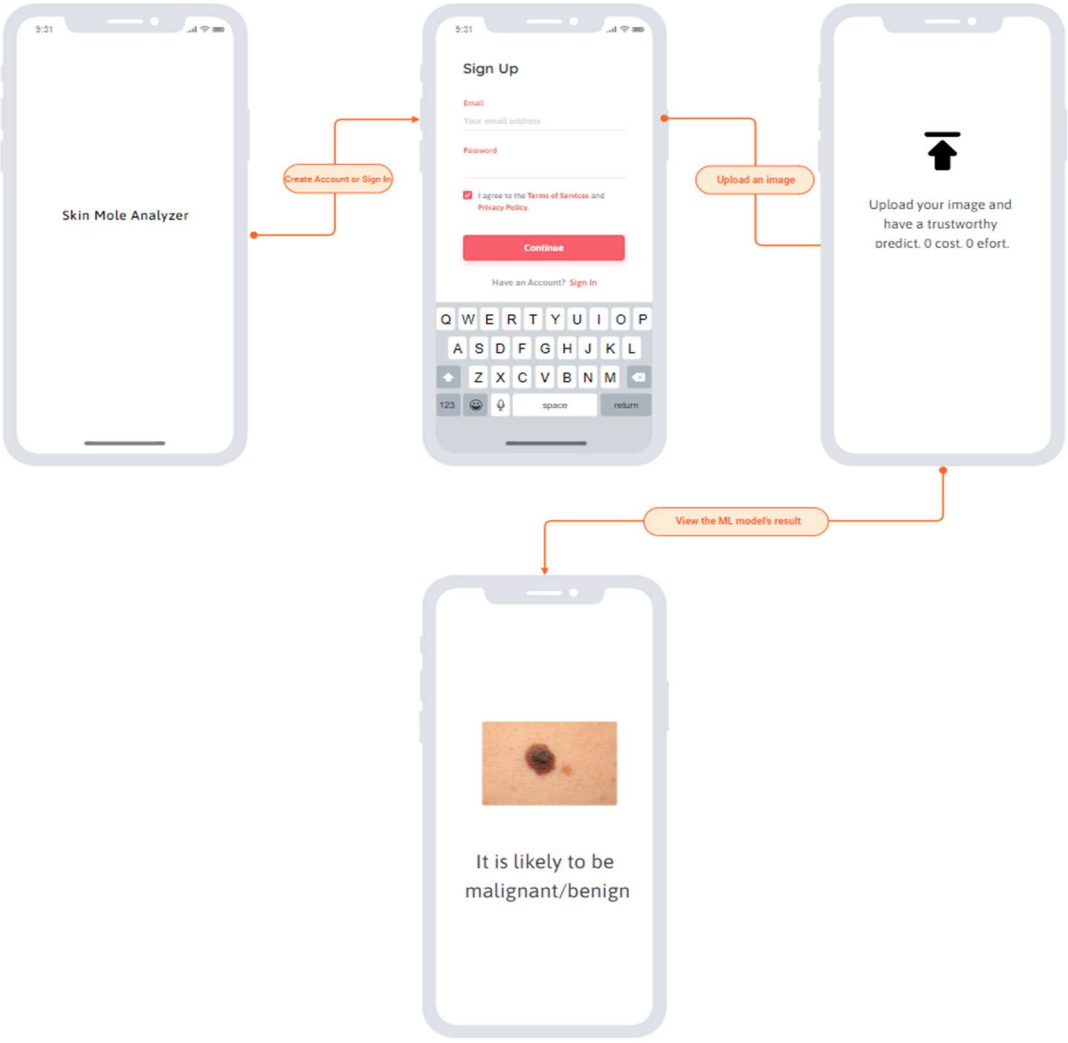


Figure 3.13 – Mockup of the application created in this project.

3.5.5 Database organization

As the user, to access the photos he tested and the results, he has to log into his account, therefore, it is necessary to structure where and how this information is stored. To do so, it is required to create a database (in this project, it was created with SQLite within the Flask framework). The structure of the database tables for this project is as follows, 1 table with the user information, a table with the information of the images that the user uploaded and finally a table with notes related to the user's images (Figure 3.14). All these tables are linked from the user id.

As in this project we are saving confidential and medical data, it is necessary to have some security regarding the database. So, a cryptographic hash function was implemented, sha-256 to be more specific. Thus, user information that is stored in the database is encrypted in order to protect confidential information.

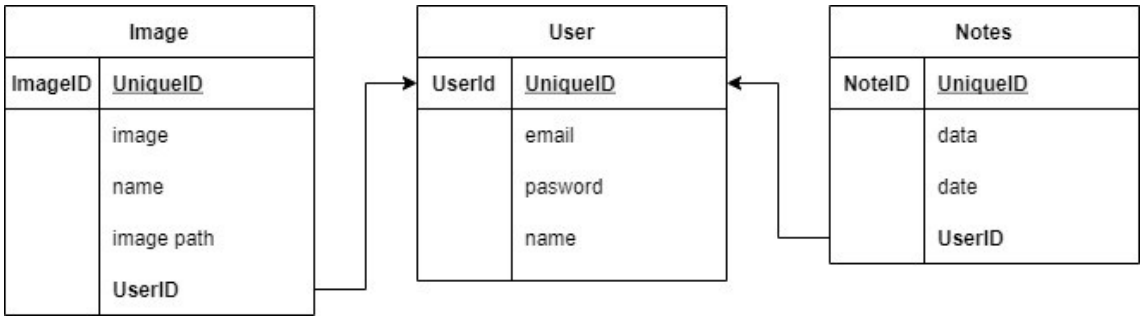


Figure 3.14 – Structure of the tables created in the database that supports the application for this project.

CHAPTER 4 – RESULTS

4. Results and Discussion

As explained before, in order to find the best model that fits the dataset created, it is going to be used the accuracy. Accuracy is the measure of all the correctly identified cases. It is generally the best measure to be used when all the classes are equally important. In this instance, the importance of the 2 classes (benign and malignant) are equally important. It is not desirable to give a benign output for malignant skin cancer, since the patient will lack early treatment, which will drastically improve the survival rates, neither it desirable the opposite, since the patient will probably be subject to medical procedures. Therefore, the model with the best accuracy overall will be considered the best model.

With this step finalized, the focus will be shifted towards the creation of a web-based application that allows the user to access the model created and quickly get the result. Unlike the Machine Learning model, for the application, there iso no measurement to be done. The requirements for the application are that it is easy and simple to use, and that, above all, it manages to allow the user to obtain a result from the ML model. To test this, and since there are no fixed measures, it was created a form where some closed relatives of mine, were prompted to answer after using the application.

Hence, there are 2 steps that define the results of this thesis:

1. Parameter selection of the Machine Learning Models
2. Mobile Application development

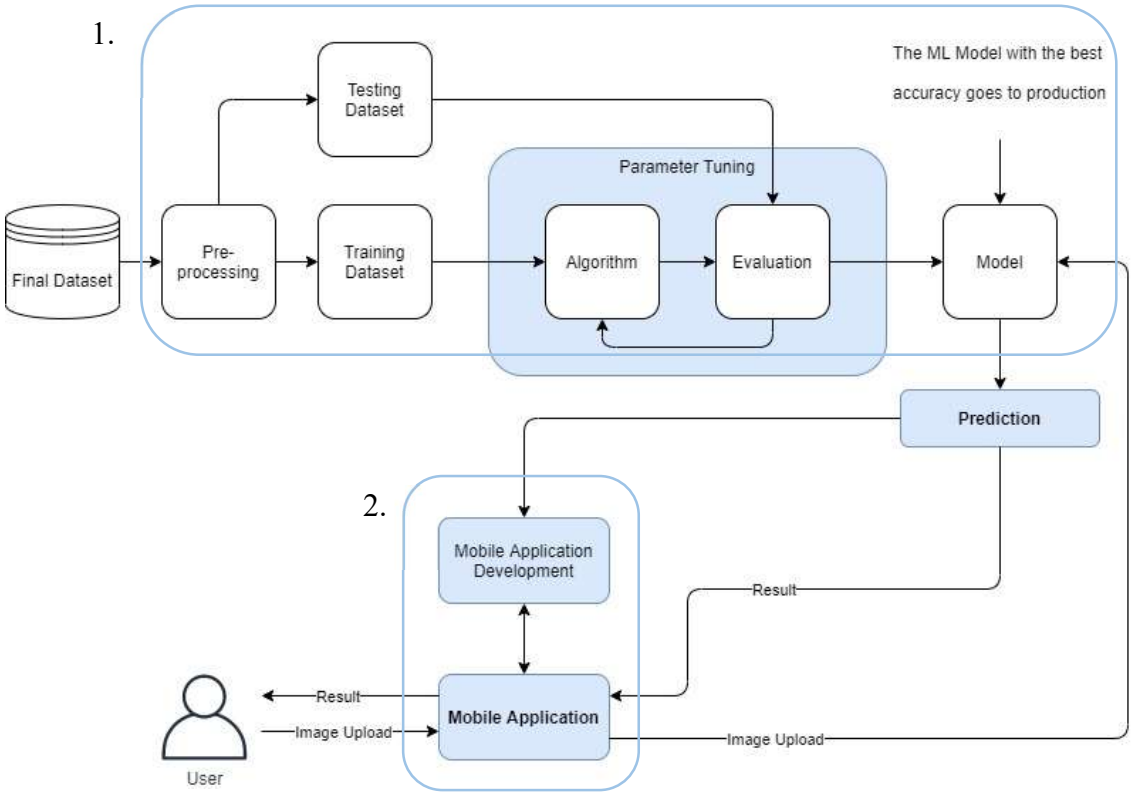


Figure 4.1 – Workflow diagram of (1.) the parameter selection for the machine learning algorithms and the (2.) mobile application development with the model with best result.

4.1 Machine Learning Models

As explained before, there were selected a few architectures to implement in the dataset created, such as VGG16, VGG19 and MobileNetV2. The 3 models were evaluated with the same measure, accuracy. Therefore, the model with the best accuracy moves forward. The results of the models can be seen below:

Table 4.1 – Accuracy of the ML models implemented in the dataset created.

ML Models	Accuracy
VGG16	91.3%
VGG19	91.7%
MobileNetV2	95.5%

4.1.1 Parameter Selection

From the analysis of ML models used in other binary classification applications, it was possible to notice that is commonly used pre-trained networks. Usually, these models have quite a success for image classification. In these models, in addition to the existing layers (described in figures 3.9, 3.10 and 3.11), new layers are typically added. In all of the models used, the weights were pre-trained on ImageNet [71, 72]. Weight is the parameter within a neural network that transforms input data within the network's hidden layers. A neural network is a series of nodes, or neurons. Within each node is a set of inputs, weight, and a bias value. Additionally, a few layers were added to provide some depth, resulting in the following:

```
Model: "VGG16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vgg16 (Functional)           (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
dense (Dense)                (None, 256)               6422784
_________________________________________________________________
dropout (Dropout)            (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 257
=================================================================
Total params: 21,137,729
Trainable params: 6,423,041
Non-trainable params: 14,714,688
_________________________________________________________________
```

Figure 4.2 – Model summary, with the addition of 2 new layers on top of the VGG16. VGG16 parameters are non-trainable since it was used a pre-trained model with fixed weights.

```
Model: "VGG19"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vgg19 (Functional)           (None, 7, 7, 512)         20024384
_________________________________________________________________
flatten_2 (Flatten)          (None, 25088)             0
_________________________________________________________________
dense_8 (Dense)              (None, 2)                 50178
=================================================================
Total params: 20,074,562
Trainable params: 50,178
Non-trainable params: 20,024,384
_________________________________________________________________
```

Figure 4.3 – Model summary, with the addition of 1 new layer on top of the VGG19. VGG19 parameters are non-trainable since it was used a pre-trained model with fixed weights.

```
Model: "MobileNetV2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
mobilenetv2_1.00_224 (Functi (None, 7, 7, 1280)        2257984
_________________________________________________________________
global_average_pooling2d (Gl (None, 1280)              0
_________________________________________________________________
dense (Dense)                (None, 320)               409920
_________________________________________________________________
dropout (Dropout)            (None, 320)               0
_________________________________________________________________
dense_1 (Dense)              (None, 320)               102720
_________________________________________________________________
dropout_1 (Dropout)          (None, 320)               0
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 642
=================================================================
Total params: 2,771,266
Trainable params: 513,282
Non-trainable params: 2,257,984
_________________________________________________________________
```

Figure 4.4 – Model summary, with the addition of 3 new layers on top of the MobileNetV2. MobileNetV2 parameters are non-trainable since it was used a pre-trained model with fixed weights.

4.1.1.1 Fine-tuning

Fine-tuning is a technique of model reusability in addition to feature extraction. Fine-tuning consists of unfreezing few of the top layers of the frozen model base in neural network used for feature extraction and jointly training both the newly added part of the model (for example, a fully connected classifier) and the top layers. This technique is called fine-tuning as it slightly adjusts the more abstract representations of fine-tuning model being reused so that it can be made more relevant for the problem at hand. Only the top layers of the convolutional base are possible to be fine-tune once the classifier on top has already been trained because to be able to train a randomly initialized classifier, freezing of pretrained convnets.

With that in mind the fine-tuning technique was applied to MobileNetV2, in order to increase the accuracy as specificity of the model and allowing the model to have better results than VGG16 and VGG19.

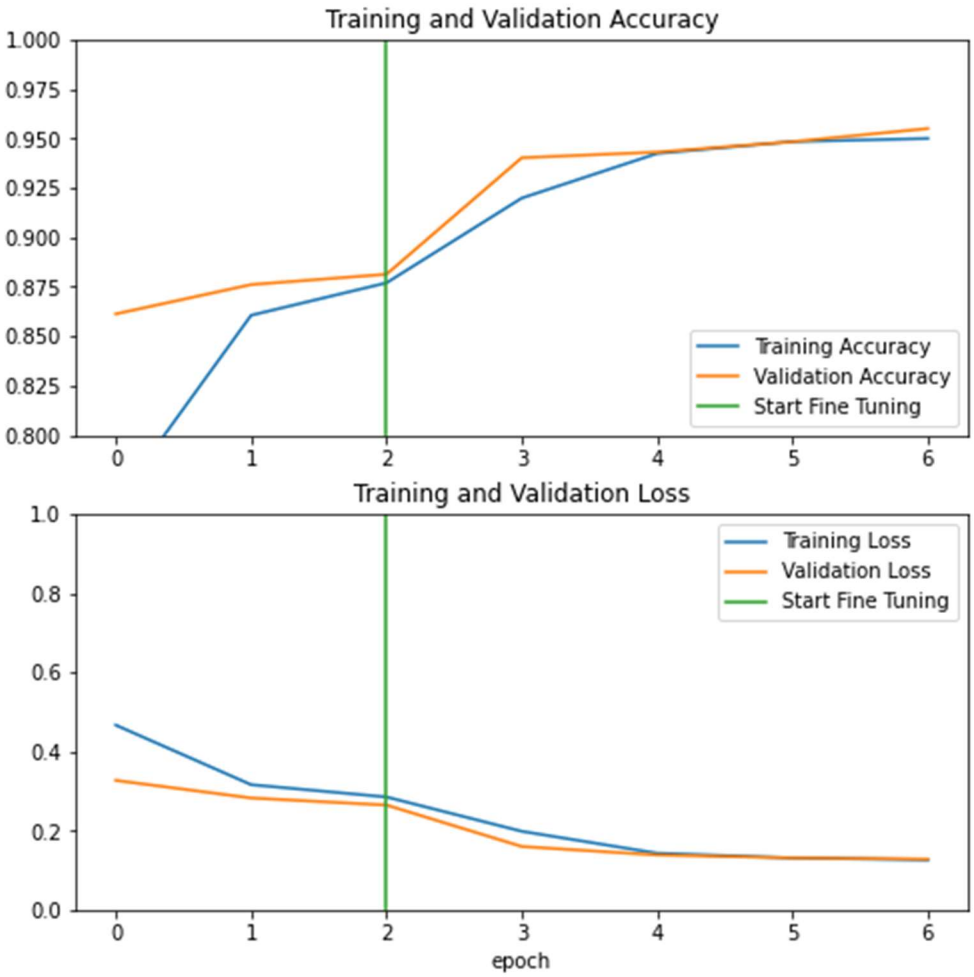


Figure 4.5 – Plot of the accuracy and loss with fine tune of MobileNetV2.

Fine-tuned, this model was able to increase the accuracy from 87,7% (with only 3 epochs) to 95,9%.

Here is of some examples of predictions of this model:

```
Predictions:
 [0 0 1 1 1 0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 1 0 0 0 0 0 0 1 1 0 0 0]
Labels:
 [0 0 0 1 1 0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 1 0 0 0 0 0 0 1 1 0 0 0]
```

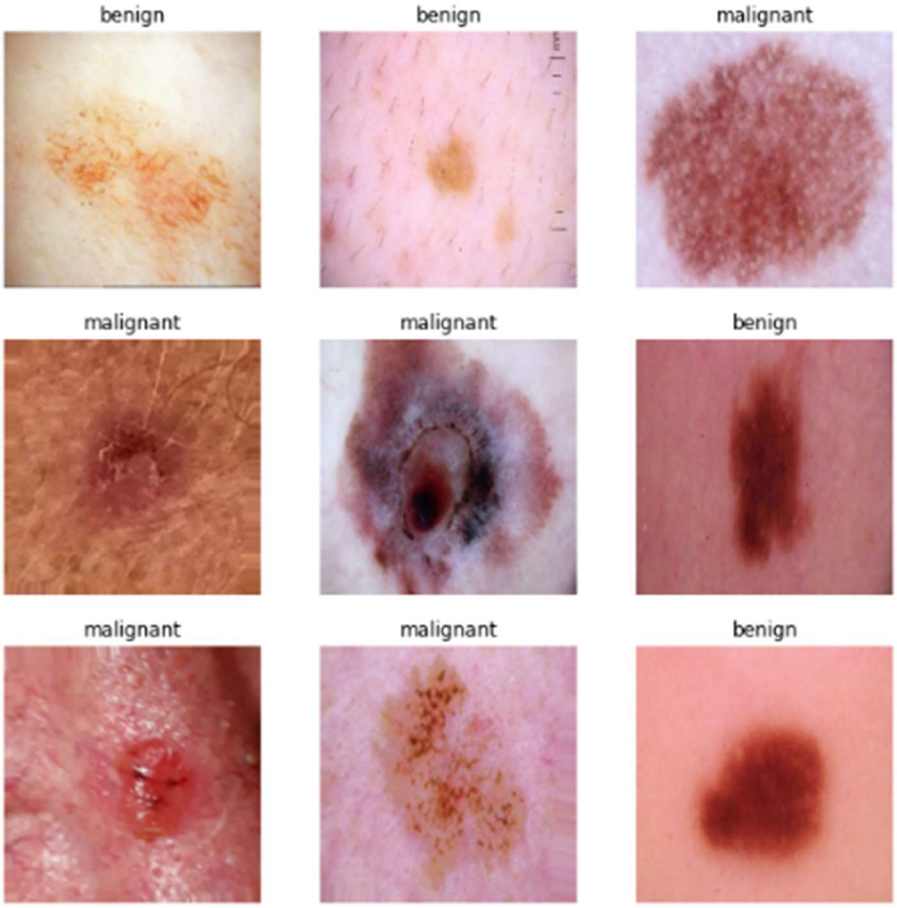


Figure 4.6 – Two arrays, one with the predictions of the model and another with the label. Below the first 9 images with the prediction on top. It is visible that in the image 3, the model predicted as “malignant”, but the label is “benign”, which categorizes this as false positive.

4.1.1.2 Fine-tuning with pre-processing

Additionally, in this thesis, it was tested if the ML model would have a better behaviour with pre-processing. Like explained before, the pre-processing used in the dataset, was a CLAHE filter (chapter 3.3.3). Comparing with the MobileNetV2 used in the original dataset, without fine-tuning, the MobileNetV2 with pre-processing had a higher accuracy with 90,6%, a difference of around 3%. However, with fine-tuning, was just a shy below the ML model with the original dataset, with 95,7% of accuracy (compared with 95,9%).

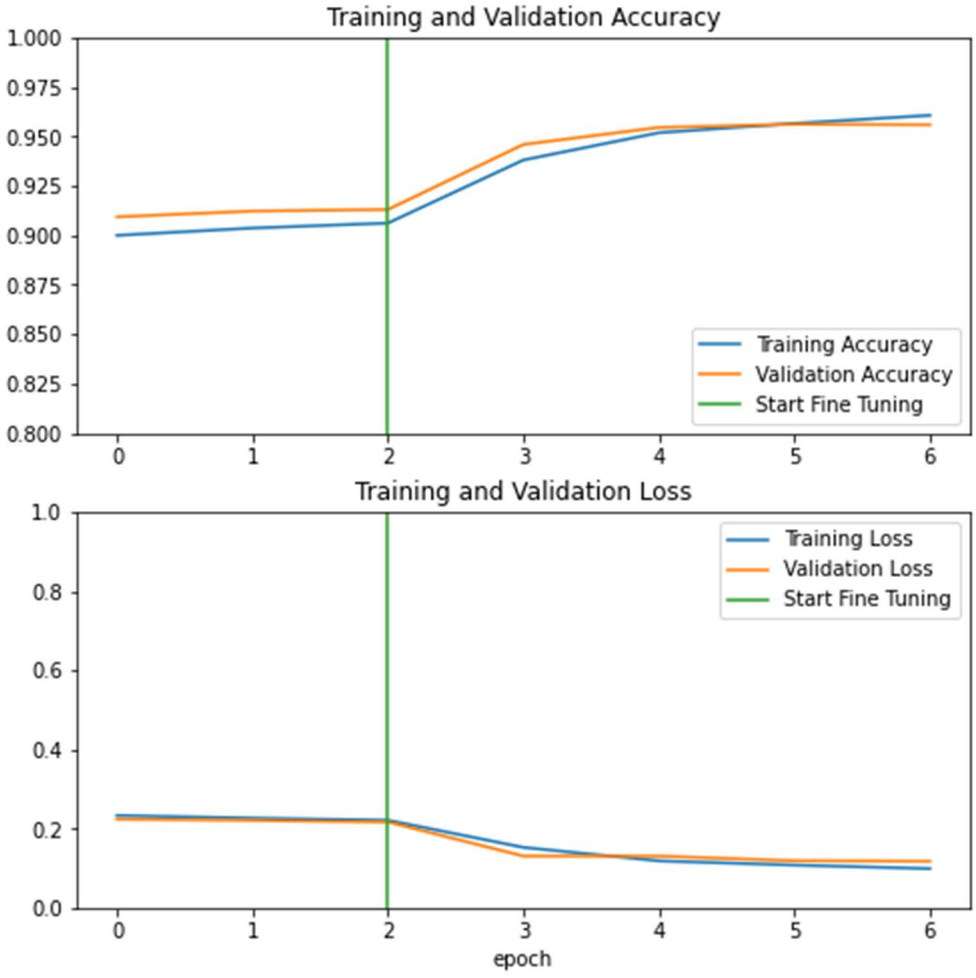


Figure 4.7 – Plot of the accuracy and loss with fine tune of MobileNetV2, with pre-processing (CLAHE filter applied in the dataset).

Here is of some examples of predictions of this model:

```
Predictions:
 [1 0 0 0 1 1 0 1 1 0 1 1 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 1 0 0 1 1]
Labels:
 [1 0 0 0 1 1 0 1 1 0 1 1 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 1 0 0 1 1]
```

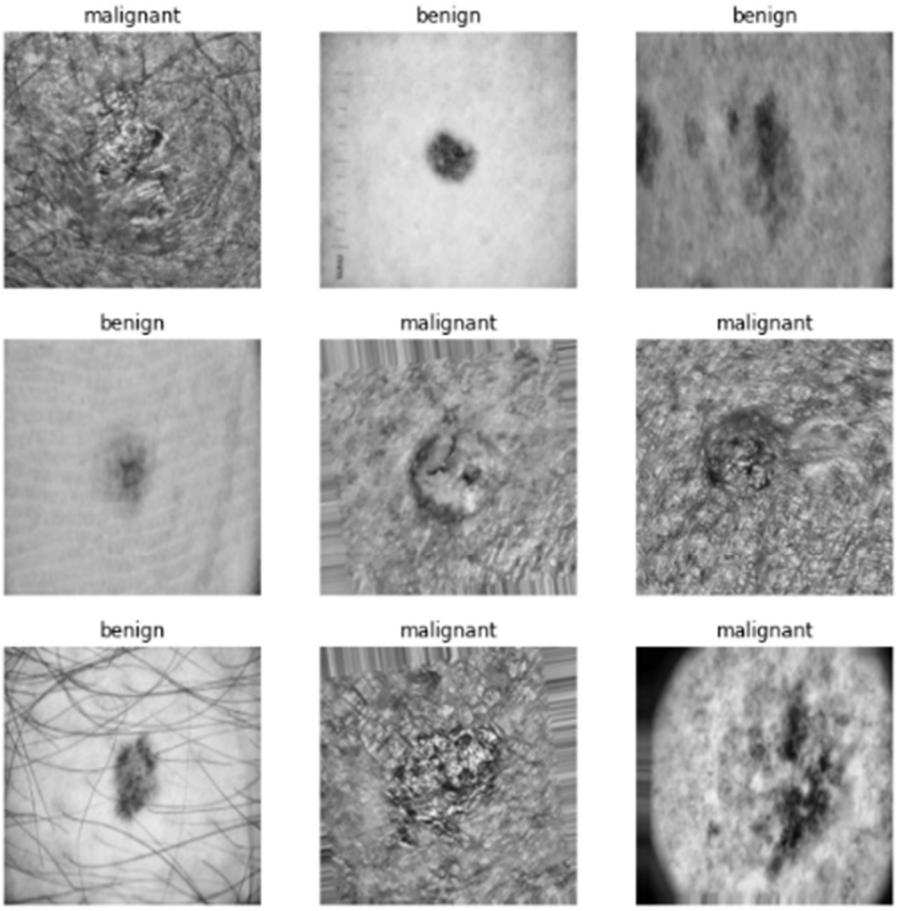


Figure 4.8 – Two arrays, one with the predictions of the model and another with the label. Below the first 9 images with the prediction on top. It is visible that in the 25th prediction of the array, the prediction the prediction was “benign” but the label is “malignant”, which categorizes this as false negative.

4.2 Application Development

The development of the application was the final part of this thesis, being just a tool that provides a usable interaction between the final user and the ML model.

Once the user accesses the application, the user must login to use it. If the account has never been created, there is a Sign-Up option on the top bar to do it.

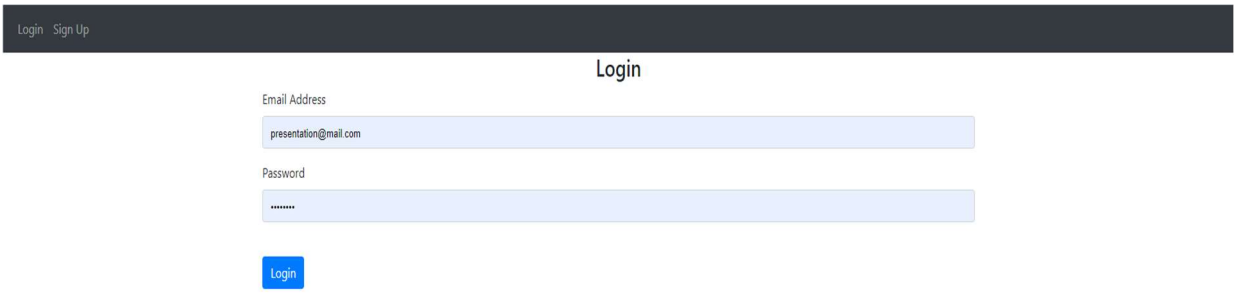


Figure 4.9 – Login page of the application created for this thesis.

When the log-in is done, it is possible to upload an image and within seconds have a result of the ML model prediction. The user can also see all the images previously upload (each user can only see the images uploaded that are associated with that account).

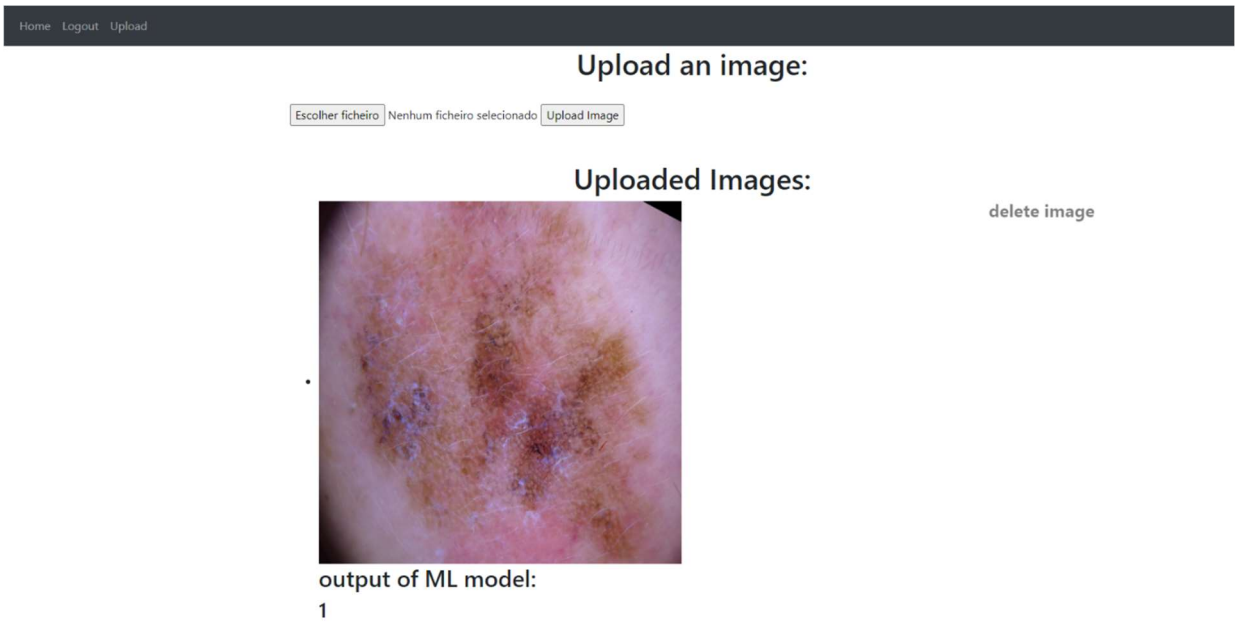


Figure 4.10 – Upload page of the application created for this thesis, with an example of an uploaded image with the result.

Since the ML model with the best result was the MobileNetV2 with the original dataset (95.9% accuracy) that was the model that is running on the backend of the application.

4.2.1 Feedback from application

Since there are no measures like in ML models, in order to understand if the application fulfills the initial requirements (simple, easy to use and provides an output from the ML model), a survey was created to have feedback on the developed application. It involved three yes or no questions:

1. Is the application intuitive?
2. Could you managed to use the application?
3. Did you understand the result?

Ten people were subjected to this survey right after they used the application, and the results for the 3 questions are the following:

1. Is the application intuitive?

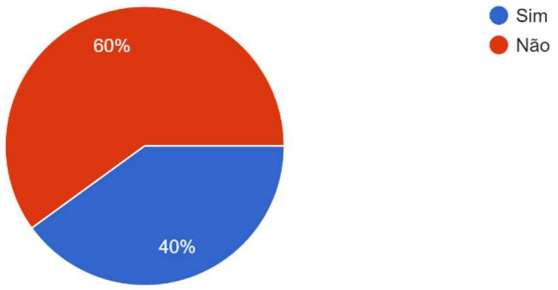


Figure 4.11 – Results for the first question of the survey. Blue for "Yes" (Sim) and red for "No" (Não).

2. Could you managed to use the application?

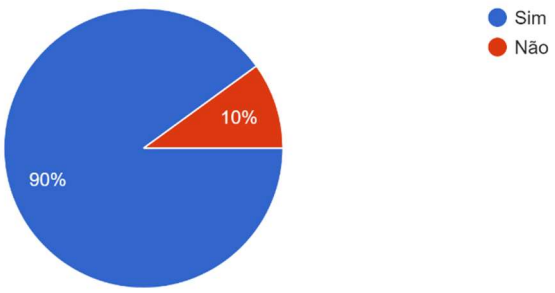


Figure 4.12 – Results for the second question of the survey. Blue for "Yes" (Sim) and red for "No" (Não).

3. Did you understand the result?

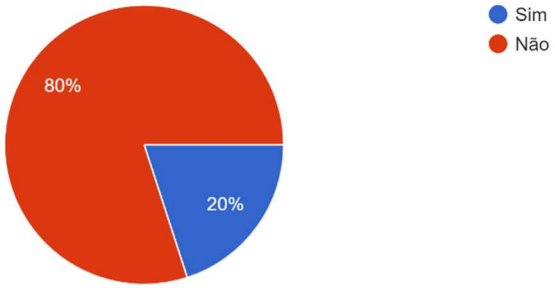


Figure 4.13 – Results for the third question of the survey. Blue for "Yes" (Sim) and red for "No" (Não).

With this it is possible to understand that almost all of the users were able to use the application (90%), however, it was neither easy or simple to both use and understand the output of the ML model.

CHAPTER 5 – CONCLUSION AND FUTURE PERSPECTIVES

5. Conclusion and Future Perspectives

5.1. Conclusion

The advances of Machine Learning and its integration into the challenges of medicine has been particularly important since it allows the creation of new tools of easy-going implementation and distribution that can make a difference in the life of patients around the world. In the last couple of years, the rates of Melanoma (and other skin cancer types) incidence and mortality have been increasing, and the projections don't induce a decrease, rather on the contrary. And since the survival rates of Melanoma and other skin cancers, is extremely higher, when it's detected in early stages, the use of applications that can through a simple cellphone provide a reliable and accurate prediction, can change the predictions of mortality by a large scale.

On this thesis, it was possible to create a complete vast dataset, to be used to train and test ML models. The dataset is a compilation of the most used public datasets, that combine more than 40000 images from patients. It was achieved a great performance on the ML model created for this dataset, with around 95% of accuracy. However, it would be excellent to have more images taken from smartphones in the dataset, since that will be the case of the users. Most of the public dataset (including most used in this thesis) are images taken with dermatoscope, and that will make a difference when we provide the trained model an image taken without the dermatoscope, since the lighting most of the times is not perfect, the distance can fluctuate a lot, and the quality of the image will depend on the camera used. Images taken with the help of the dermatoscope will generally have the same patters almost every time.

The mobile application was not as successful as the ML model, since it only succeeded in one pilar of the three that were supposed in the beginning. Nevertheless, the application was never the focus of this thesis. It was looked as a tool to provide easier access of the ML model to the final user. With that in mind, to overcome these difficulties, it would be necessary to engage more time and effort on its development, which was used in the ML model.

The ML model, with an exceptionally high accuracy, will certainly help on the development and research of new methods to overcome cancer mortality and incidence rates.

5.2. Future Perspectives

- What if I had more "normal" images in my dataset?

That is the main question raised with the results presented. It would certainly be a progress over what was achieved in this project. The lack of images taken without the help of a dermatoscope will cause a lot of wrong predictions in a large-scale tests, since the images will be completely different from the majority of images used to train the model.

- What if I had more processing power to develop the ML model?

The processing power would probably not do much on the MobileNet V2 since it's a ML model with an architecture created to be used with low processing power. However, that cannot be said for the VGG16 and VGG19, which would probably be easier to tune and train with more processing power and time, which could lead to better results on those architectures. It would also be useful to implement more types of architectures, not only three.

- Would I need to invest some more time on the web-based application?

The mobile application is for sure on the backbones of this project, thus it would be mandatory to improve its usability and scalability for a real usage scenario. The design of the pages should be taken in consideration as well as the protection and security of the patient's data, since it involves medical information.

References

[1] J. Daghrir, L. Tlig, M. Bouchouicha, and M. Sayadi, "Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach," *2020 Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2020*, Sep. 2020, doi: 10.1109/ATSIP49331.2020.9231544.

[2] B. Bandarchi, L. Ma, R. Navab, A. Seth, and G. Rasty, "From melanocyte to metastatic malignant melanoma," *Dermatol. Res. Pract.*, vol. 2010, no. 1, 2010, doi: 10.1155/2010/583748.

[3] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[4] M. Gniadecka *et al.*, "Melanoma Diagnosis by Raman Spectroscopy and Neural Networks: Structure Alterations in Proteins and Lipids in Intact Cancer Tissue," *J. Invest. Dermatol.*, vol. 122, no. 2, pp. 443–449, 2004, doi: 10.1046/j.0022-202X.2004.22208.x.

[5] H. A. Haenssle *et al.*, "Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018, doi: 10.1093/annonc/mdy166.

[6] R. A. Scolyer, G. V. Long, and J. F. Thompson, "Evolving concepts in melanoma classification and their relevance to multidisciplinary melanoma patient care," *Mol. Oncol.*, vol. 5, no. 2, pp. 124–136, 2011, doi: 10.1016/j.molonc.2011.03.002.

[7] Y. Liu and M. S. Sheikh, "Melanoma: Molecular pathogenesis and therapeutic management," *Mol. Cell. Pharmacol.*, vol. 6, no. 3, pp. 31–44, 2014, doi: 10.4255/mcpharmacol.14.03.

[8] "Cutaneous Melanoma | Melanoma Research Foundation." https://melanoma.org/patients-caregivers/cutaneous-melanoma/ (accessed Mar. 29, 2021).

[9] "Melanoma: Practice Essentials, Overview, Indications and Guidelines." https://emedicine.medscape.com/article/1295718-overview#aw2aab6b2 (accessed Mar. 29, 2021).

[10] Y. Wang, Y. Zhao, and S. Ma, "Racial differences in six major subtypes of melanoma: Descriptive epidemiology," *BMC Cancer*, vol. 16, no. 1, 2016, doi: 10.1186/s12885-016-2747-6.

[11] A. Sboner *et al.*, "A multiple classifier system for early melanoma diagnosis," *Artif. Intell. Med.*, vol. 27, no. 1, pp. 29–44, 2003, doi: 10.1016/S0933-3657(02)00087-8.

[12] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *Lancet Oncol.*, vol. 3, no. 3, pp. 159–165, 2002, doi: 10.1016/S1470-2045(02)00679-4.

[13] K. Westerhoff, W. H. McCarthy, and S. W. Menzies, "Increase in the sensitivity for melanoma diagnosis by primary care physicians using skin surface microscopy," *Br. J. Dermatol.*, vol. 143, no. 5, pp. 1016–1020, 2000, doi: 10.1046/j.1365-2133.2000.03836.x.

[14] I. R. Bristow and J. Bowling, "Dermoscopy as a technique for the early identification of foot melanoma," *J. Foot Ankle Res.*, vol. 2, no. 1, 2009, doi: 10.1186/1757-1146-2-14.

[15] "ISIC Archive." https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main (accessed

Aug. 15, 2021).

[16] "Lentigo maligna and lentigo maligna melanoma dermoscopy | DermNet NZ." https://dermnetnz.org/topics/lentigo-maligna-and-lentigo-maligna-melanoma-dermoscopy/ (accessed Mar. 31, 2021).

[17] Y. Teramoto, H. Martinez-Said, J. Guo, and C. Garbe, "Acral Lentiginous Melanoma," in *Cutaneous Melanoma*, Cham: Springer International Publishing, 2019, pp. 1–28.

[18] D. S. Rigel, R. J. Friedman, A. W. Kopf, and D. Polsky, "ABCDE - An evolving concept in the early detection of melanoma," *Arch. Dermatol.*, vol. 141, no. 8, pp. 1032–1034, 2005, doi: 10.1001/archderm.141.8.1032.

[19] E. Real *et al.*, "Depth-resolved attenuation coefficient estimation for skin cancer assessment with Optical Coherence Tomography," *Clin. Preclin. Opt. Diagnostics II (2019), Pap. 11078_37*, vol. Part F142-ECBO 2019, p. 11078_37, Jun. 2019, doi: 10.1117/12.2527255.

[20] V. W. Rebecca, V. K. Sondak, and K. S. M. Smalley, "A brief history of melanoma," *Melanoma Res.*, vol. 22, no. 2, pp. 114–122, 2012, doi: 10.1097/cmr.0b013e328351fa4d.

[21] Breslow A, "Thickness, Cross-Sectional Areas and Depth of Invasion in the Prognosis of Cutaneous Melanoma," *Annals of Surgery*, vol. 172, no. 5. pp. 902–908, 1970.

[22] P. V. Dickson and J. E. Gershenwald, "Staging and prognosis of cutaneous melanoma," *Surgical Oncology Clinics of North America*, vol. 20, no. 1. Surg Oncol Clin N Am, pp. 1–17, Jan. 2011, doi: 10.1016/j.soc.2010.09.007.

[23] E. Z. Keung and J. E. Gershenwald, "The eighth edition American Joint Committee on Cancer (AJCC) melanoma staging system: implications for melanoma treatment and care," *Expert Review of Anticancer Therapy*, vol. 18, no. 8. Taylor and Francis Ltd, pp. 775–784, Aug. 03, 2018, doi: 10.1080/14737140.2018.1489246.

[24] F. Dimitriou *et al.*, "The World of Melanoma: Epidemiologic, Genetic, and Anatomic Differences of Melanoma Across the Globe," *Curr. Oncol. Rep.*, vol. 20, no. 11, 2018, doi: 10.1007/s11912-018-0732-8.

[25] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[26] M. Krensel, I. Schäfer, and M. Augustin, "Cost-of-illness of melanoma in Europe – a modelling approach," *J. Eur. Acad. Dermatology Venereol.*, vol. 33, pp. 34–45, 2019, doi: 10.1111/jdv.15308.

[27] E. Wee, R. Wolfe, C. Mclean, J. W. Kelly, and Y. Pan, "The anatomic distribution of cutaneous melanoma: A detailed study of 5141 lesions," *Australas. J. Dermatol.*, vol. 61, no. 2, pp. 125–133, May 2020, doi: 10.1111/ajd.13223.

[28] L. Deng and D. Yu, "Deep Learning: Methods and Applications," *Found. Trends® Signal Process.*, vol. 7, no. 3–4, pp. 197–387, 2014, doi: 10.1561/2000000039.

[29] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, “Medical image classification with convolutional neural network,” *2014 13th Int. Conf. Control Autom. Robot. Vision, ICARCV 2014*, vol. 2014, no. December, pp. 844–848, 2014, doi: 10.1109/ICARCV.2014.7064414.

[30] L. G. Valiant, “A theory of the learnable,” *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, Nov. 1984, doi: 10.1145/1968.1972.

[31] C. Robert, “Machine Learning, a Probabilistic Perspective,” *https://doi.org/10.1080/09332480.2014.914768*, vol. 27, no. 2, pp. 62–63, Apr. 2014, doi: 10.1080/09332480.2014.914768.

[32] B. Devore-Mcdonald and E. D. Berger, “Mossad: Defeating Software Plagiarism Detection,” 2020, doi: 10.1145/0000000.0000000.

[33] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, “An Overview of Machine Learning,” *Mach. Learn.*, pp. 3–23, 1983, doi: 10.1007/978-3-662-12405-5_1.

[34] J. A. Adamu, “Advanced Stochastic Optimization Algorithm for Deep Learning Artificial Neural Networks in Banking and Finance Industries,” *Risk Financ. Manag.*, vol. 1, no. 1, p. p8, Nov. 2019, doi: 10.30560/RFM.V1N1P8.

[35] D. J. Norris, “Beginning Artificial Intelligence with the Raspberry Pi,” *Begin. Artif. Intell. with Raspberry Pi*, 2017, doi: 10.1007/978-1-4842-2743-5.

[36] J. Schmidhuber, “Deep Learning in Neural Networks: An Overview,” *Neural Networks*, vol. 61, pp. 85–117, Apr. 2014, doi: 10.1016/j.neunet.2014.09.003.

[37] I. Castiglioni *et al.*, “AI applications to medical images: From machine learning to deep learning,” *Phys. Medica Eur. J. Med. Phys.*, vol. 83, pp. 9–24, Mar. 2021, doi: 10.1016/J.EJMP.2021.02.006.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Accessed: Oct. 25, 2021. [Online]. Available: http://code.google.com/p/cuda-convnet/.

[39] C. Szegedy *et al.*, “Going deeper with convolutions,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June-2015, pp. 1–9, Oct. 2015, doi: 10.1109/CVPR.2015.7298594.

[40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2016.

[41] W. Setiawan and F. Damayanti, “Layers Modification of Convolutional Neural Network for Pneumonia Detection,” *J. Phys. Conf. Ser.*, vol. 1477, no. 5, 2020, doi: 10.1088/1742-6596/1477/5/052055.

[42] N. Jankowski and M. Grochowski, “Comparison of Instances Seletion Algorithms I. Algorithms Survey,” *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.*, vol. 3070, pp. 598–603, 2004, doi: 10.1007/978-3-540-24844-6_90.

[43] A. Hidaka and T. Kurita, “Consecutive Dimensionality Reduction by Canonical Correlation Analysis for Visualization of Convolutional Neural Networks,” *Proc. Isc. Int. Symp. Stoch. Syst.*

Theory its Appl., vol. 2017, no. 0, pp. 160–167, 2017, doi: 10.5687/SSS.2017.160.

[44] J. Latif, C. Xiao, A. Imran, and S. Tu, "Medical imaging using machine learning and deep learning algorithms: A review," *2019 2nd Int. Conf. Comput. Math. Eng. Technol. iCoMET 2019*, Mar. 2019, doi: 10.1109/ICOMET.2019.8673502.

[45] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran, "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," *Proc. - Int. Conf. Pattern Recognit.*, vol. 0, pp. 1195–1200, Jan. 2016, doi: 10.1109/ICPR.2016.7899799.

[46] E. Kim, M. Corte-Real, and Z. Baloch, "A deep semantic mobile application for thyroid cytopathology," *Med. Imaging 2016 PACS Imaging Informatics Next Gener. Innov.*, vol. 9789, p. 97890A, Apr. 2016, doi: 10.1117/12.2216468.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June-2015, pp. 431–440, Oct. 2015, doi: 10.1109/CVPR.2015.7298965.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[49] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9901 LNCS, pp. 424–432, Oct. 2016, doi: 10.1007/978-3-319-46723-8_49.

[50] "The PASCAL Visual Object Classes Challenge 2012 (VOC2012)." http://host.robots.ox.ac.uk/pascal/VOC/voc2012/ (accessed Oct. 25, 2021).

[51] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A Survey on Deep Transfer Learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11141 LNCS, pp. 270–279, Oct. 2018, doi: 10.1007/978-3-030-01424-7_27.

[52] R. Iniesta, D. Stahl, and P. McGuffin, "Machine learning, statistical learning and the future of biological research in psychiatry," *Psychological Medicine*, vol. 46, no. 12. Cambridge University Press, pp. 2455–2465, Sep. 01, 2016, doi: 10.1017/S0033291716001367.

[53] H. TL, L. WR, H. SC, and F. JY, "Cisplatin encapsulated in phosphatidylethanolamine liposomes enhances the in vitro cytotoxicity and in vivo intratumor drug accumulation against melanomas," *J. Dermatol. Sci.*, vol. 46, no. 1, pp. 11–20, Apr. 2007, doi: 10.1016/J.JDERMSCI.2006.12.011.

[54] U. Jamil, A. Sajid, M. Hussain, O. Aldabbas, A. Alam, and M. U. Shafiq, "Melanoma segmentation using bio-medical image analysis for smarter mobile healthcare," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 10, pp. 4099–4120, Oct. 2019, doi: 10.1007/S12652-019-01218-0.

[55] N. Cheerla and D. Frazier, "Automatic Melanoma Detection Using Multi-Stage Neural Networks," *Int. J. Innov. Res. Sci. Eng. Technol. (An ISO*, vol. 3297, no. 2, pp. 2319–8753, 2007, Accessed:

Jul. 28, 2021. [Online]. Available: www.ijirset.com.

[56] M. Ruela, C. Barata, T. Mendonca, and J. S. Marques, “On the role of shape in the detection of melanomas,” *Int. Symp. Image Signal Process. Anal. ISPA*, pp. 268–273, 2013, doi: 10.1109/ISPA.2013.6703751.

[57] J. Daghrir, L. Tlig, M. Bouchouicha, and M. Sayadi, “Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach,” *2020 Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2020*, Sep. 2020, doi: 10.1109/ATSIP49331.2020.9231544.

[58] Q. Ha, B. Liu, and F. Liu, “Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge,” 2020, Accessed: Jul. 28, 2021. [Online]. Available: https://github.com/haqishen/SIIM-ISIC-Melanoma-Classification-1st-Place-Solution.

[59] “SIIM-ISIC Melanoma Classification | Kaggle.” https://www.kaggle.com/c/siim-isic-melanoma-classification (accessed Aug. 15, 2021).

[60] P. Tschandl, C. Rosendahl, and H. Kittler, “Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci. Data*, vol. 5, pp. 1–9, 2018, doi: 10.1038/sdata.2018.161.

[61] T.-T. Do *et al.*, “Accessible Melanoma Detection using Smartphones and Mobile Image Analysis,” Accessed: Aug. 16, 2021. [Online]. Available: www.dermoscopy.org,.

[62] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, “PH2 - A dermoscopic image database for research and benchmarking,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 5437–5440, 2013, doi: 10.1109/EMBC.2013.6610779.

[63] N. Japkowicz, N. Japkowicz, and S. Stephen, “The class imbalance problem: A systematic study,” *Intell. DATA Anal.*, pp. 429--449, 2002, Accessed: Aug. 17, 2021. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.711.8214.

[64] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. J. Aguirre, and A. M. Vanegas, “Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models,” *Sensors 2020, Vol. 20, Page 4373*, vol. 20, no. 16, p. 4373, Aug. 2020, doi: 10.3390/S20164373.

[65] R. Roslidar, K. Saddami, F. Arnia, M. Syukri, and K. Munadi, “A study of fine-tuning CNN models based on thermal imaging for breast cancer classification,” *Proc. Cybern. 2019 - 2019 IEEE Int. Conf. Cybern. Comput. Intell. Towar. a Smart Human-Centered Cyber World*, pp. 77–81, Aug. 2019, doi: 10.1109/CYBERNETICSCOM.2019.8875661.

[66] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, Sep. 2014, Accessed: Aug. 17, 2021. [Online]. Available: https://arxiv.org/abs/1409.1556v6.

[67] H. Qassim, A. Verma, and D. Feinzimer, “Compressed residual-VGG16 CNN model for big data places image recognition,” *2018 IEEE 8th Annu. Comput. Commun. Work. Conf. CCWC 2018*, vol.

2018-January, pp. 169–175, Feb. 2018, doi: 10.1109/CCWC.2018.8301729.

[68] "A Gentle Introduction to Pooling Layers for Convolutional Neural Networks." https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/ (accessed Aug. 17, 2021).

[69] Y. Zheng, C. Yang, and A. Merkulov, "Breast cancer screening using convolutional neural network and follow-up digital mammography," p. 4, May 2018, doi: 10.1117/12.2304564.

[70] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4510–4520, Jan. 2018, Accessed: Aug. 17, 2021. [Online]. Available: https://arxiv.org/abs/1801.04381v4.

[71] A. Jha, M. Dave, and S. Madan, "Comparison of Binary Class and Multi-Class Classifier Using Different Data Mining Classification Techniques," pp. 13–14, 2019, Accessed: Aug. 18, 2021. [Online]. Available: https://ssrn.com/abstract=3464211.

[72] "Welcome to Flask — Flask Documentation (2.0.x)." https://flask.palletsprojects.com/en/2.0.x/ (accessed Aug. 21, 2021).

[73] S. Liawatimena *et al.*, "Django Web Framework Software Metrics Measurement Using Radon and Pylint," *1st 2018 Indones. Assoc. Pattern Recognit. Int. Conf. Ina. 2018 - Proc.*, pp. 218–222, Jan. 2019, doi: 10.1109/INAPR.2018.8627009.

[74] "MobileNet, MobileNetV2, and MobileNetV3." https://keras.io/api/applications/mobilenet/ (accessed Oct. 09, 2022).

[75] "VGG16 and VGG19." https://keras.io/api/applications/vgg/ (accessed Oct. 09, 2022).