

Faculty of Technology

School of Computer Science

***PROM06 – MSc Dissertation***

***MSc Computer Science with Data Science***

Student Name: Giulio Leopaldi

Student Number: bh87ef

Supervisor Name: Prof Ken McGarry

Improving Melanoma Detection with   
MobileNet V3:   
A Comparison to MobileNet V2 and Exploration of Transfer Learning and Fine-Tuning Techniques

March 2023

**Declaration**

I declare the following:

(1) that the material contained in this dissertation is the end result of my own work and that due acknowledgement has been given in the bibliography and references to **ALL** sources be they printed, electronic or personal.

(2) the Word Count of this Dissertation is 10,926.

(3) that unless this dissertation has been confirmed as confidential, I agree to an entire electronic copy or sections of the dissertation being placed on the eLearning Portal, if deemed appropriate, to allow future students the opportunity to see examples of past dissertations.  I understand that if displayed on the eLearning Portal it would be made available for no longer than five years and that students would be able to print off copies or download.

(4) I agree to my dissertation being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other Department or from other institutions using the service.

In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and second marker, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.

(5) I have read the University of Sunderland Policy Statement on Ethics in Research and I confirm that ethical issues have been considered, evaluated and appropriately addressed in this research.

**SIGNED:**

**DATE: 2nd March 2023**

*To my wife Frieda, thank you for being my rock throughout this journey. Your unwavering support, patience, and encouragement kept me going during the challenging moments of this journey.*

*To all my family who always believed I could do it and supported me unconditionally.*

**Abstract**

Early detection of melanoma, the deadliest form of skin cancer, increases patient survival rates. Artificial intelligence software that automates melanoma diagnosis can aid clinical practices and potentially save lives. Artificial neural networks are a deep learning technique that processes information similarly to the human brain. The popularity of smartphone applications with embedded lightweight neural networks is on the rise, offering numerous opportunities to expand research in this field.

This study firstly compares the performance of melanoma detection models on the HAM 10000 dataset based on transfer learning from MobileNet v3-Small and MobileNet v2. The results show that the base model MobileNet v3-Small outperforms MobileNet v2 in terms of accuracy, sensitivity, specificity, and computational timing for the training and for the predictions.

Ultimately, a fine-tuned model is created on top of the base model, achieving an accuracy of 76%, and specificity and sensitivity both of 94%. This result is obtained by leveraging transfer learning from the MobileNet v3-Small and adding five fully connected layers on top of the pre-trained model to capture the specific features for this task. This is further optimised by applying data augmentation and hyper-parameter tuning methods.  
Whilst this study has limitations, including a relatively small dataset, potential underrepresentation of certain groups, and evaluation of the model's performance on a single device, it nonetheless provides promising results that can guide future research in this area.

**Contents**

[1 Introduction 6](#_Toc128864299)

[1.1 Background 6](#_Toc128864300)

[1.2 Aims and Objectives 6](#_Toc128864302)

[1.3 Research Approach and Structure of the Report 6](#_Toc128864306)

[2 Literature Review 8](#_Toc128864307)

[2.1 Contextualising the Research Problem 8](#_Toc128864308)

[2.2 Machine Learning 9](#_Toc128864309)

[2.3 Artificial Neural Networks and Deep Learning 11](#_Toc128864310)

[2.4 Convolutional Neural Networks 12](#_Toc128864311)

[2.5 MobileNet models 15](#_Toc128864312)

[3 Practical Research Methodology 17](#_Toc128864313)

[3.1 Definition of requirements 17](#_Toc128864314)

[3.1.1 Data requirements 17](#_Toc128864315)

[3.1.2 Ethical approval requirements 17](#_Toc128864316)

[3.1.3 Hardware and software requirements 17](#_Toc128864317)

[3.1.4 Pre-processing requirements 19](#_Toc128864318)

[3.1.7 Evaluation metrics requirements 19](#_Toc128864319)

[3.2 Experiment design 20](#_Toc128864320)

[3.2.1 Step 1: downloading and pre-processing the data 20](#_Toc128864321)

[3.2.2 Step 2 - The base model 22](#_Toc128864322)

[3.2.3 Step 3 - The comparison model 23](#_Toc128864323)

[3.2.4 Step 4 - Fine-tuning the base model 23](#_Toc128864324)

[3.3 Collection of results 25](#_Toc128864325)

[4 Results: Analysis and Evaluation 27](#_Toc128864326)

[4.1 Analysis of results 27](#_Toc128864327)

[4.2 Discussion of results 28](#_Toc128864328)

[4.3 Limitations of this study 30](#_Toc128864329)

[4.4 Future Research 31](#_Toc128864330)

[5 Project Evaluation and Reflection 32](#_Toc128864331)

[5.1 Overview of Project and Objectives Review 32](#_Toc128864332)

[5.2 Evaluation of methodology and results 32](#_Toc128864333)

[5.3 Alternative aims or hypotheses 33](#_Toc128864334)

[5.4 Evaluation of Ethical, Legal, Social, Security and Professional Considerations 34](#_Toc128864335)

[5.5 Personal reflection 35](#_Toc128864336)

[6 Conclusion and Recommendations 37](#_Toc128864337)

[7 Reference List 39](#_Toc128864338)

Appendix A. Research Proposal 43

Appendix B. Code for this project 44

Appendix C. Artifacts in Dermoscopy images 45

Appendix D. Skin lesion images 46

Appendix E. Artificial Neural Network architecture representation………………………..47

Appendix F. Convolutional Neural Network representation 48

Appendix G. Testing set predictions: results and confusion matrix 49

# Introduction

## Background

## Melanoma (ME) is the most fatal form of skin cancer, but early detection during its development stages increases the rates of patient survival (Melanoma skin cancer, 2022). In order to help with early diagnosis, artificial intelligence (AI) software that utilise artificial neural networks (ANN) have been developed to automate the diagnosis of ME, which can help save lives. The field of ANN has made significant advancements in recent years, with a trend towards developing practical and versatile lightweight models, rather than focusing on deep learning models that offer maximum statistical performance but are computationally expensive. This has allowed for more efficient and effective use of AI in clinical practice for ME diagnosis (Udrea et al., 2019).

## Aims and Objectives

## The purpose of this project is to advance research in the field by exploring new applications for detecting ME. Given the growing demand for healthcare apps that aid clinical practices, lightweight models that can be embedded in smartphones are rising to facilitate early detection of this disease (Rat et al., 2018). An in-depth review of existing literature in Chapter 2 has been carried out to identify the latest tendencies and areas where research is lacking. The identified trend towards lightweight models for ME detection has determined the objectives of this project, which are framed as the following questions:

## How does a MobileNet v3-Small (MNv3-S) architecture compare to a MobileNet v2 (MNv2) in terms of accuracy, sensitivity, specificity, and computational timing?

## What level of performance can be achieved through fine-tuning the MNv3-S base model in terms of accuracy, sensitivity, and specificity?

## Research Approach and Structure of the Report

To accomplish these objectives, a thorough experimentation plan has been designed, consisting of four clear and distinct steps, described in detail in Chapter 3: data acquisition and pre-processing, base model creation using MNv3-S, comparison model creation using MNv2, and fine-tuning of the final model built upon the base model.   
Chapter 4 presents the findings of the study, offering an analysis and interpretation of the results whilst also outlining the limitations of the research.  
A comprehensive evaluation of this study is presented in chapter 5 by reviewing the objectives and assessing the validity, reliability, robustness, and reproducibility of this project. Additionally, section 5.4 explores important ethical, legal, social, and considerations that must be taken into account.  
A summary, conclusions and future directions following this project can be found in the final section of this paper, in chapter 6.

# Literature Review

## Contextualising the Research Problem

ME is the most dangerous kind of skin cancers and, generally, one of the most common types of cancers. Concerning figures have shown a growth of its incidence each year (The International Agency for Research on Cancer (IARC)). Over the past decade, there has been a 32% increase in ME incidence rates in the United Kingdom. It is now the fifth most common type of cancer, accounting for 4% of all new cancer cases, and leading to approximately 2,300 deaths annually. These figures are estimated to keep growing by another 7% by 2035, reaching a number of cases of 32 per 100,000 people. However, there is some positive news in that mortality rates are projected to decrease. This is due in part to the raised awareness surrounding this disease, as well as advancements in technology that aid in early diagnosis. Prevention also plays a crucial role, as it is estimated that 86% of ME skin cancer cases are caused by behavioural and lifestyle factors. One of the highest risk factors for developing ME is posed by sunburn and the exposure to ultraviolet (UV) rays which come from the sun or artificial sources such as tanning beds or sunlamps. To reduce ME risk, health guidelines recommend avoiding direct sunlight exposure during peak hours, particularly in the summer. This can be achieved by staying in the shade, wearing appropriate clothing that covers the arms and legs, and using sunscreen with a high sun protection factor (SPF). It is also recommended to avoid indoor, artificial tanning, as this also exposes the skin to harmful UV rays (Leopaldi, 2022).   
ME often develops as an abnormal growth of melanocyte skin cells that produce and contain a pigment called melanin, responsible for the protective effect of the darkening of skin following sun exposure. It also determines the skin, hair and eyes colour. Skin lesions, areas that have distinct features from the surrounding skin, are very common and usually do not pose a threat. A clustering of melanocytes can generate moles (nevi), which are a common type of skin growth and appear as small, dark spots on the surface of the skin. Most often they result as benign, therefore harmless. However, some moles can be a sign of malignant ME, the most dangerous type of skin cancer. The primary treatment is surgical removal of the tumour which can significantly increase the chances of a positive outcome, with a 5-year survival rate of 99% (Melanoma skin cancer, 2022). However, a delayed diagnosis of ME can have significant negative impacts on an individual's health and life as it can spread systemically to other parts of the body and organs, leading to a more complicated treatment process and higher mortality rates (Leopaldi, 2022).   
Delayed diagnosis generally occurs due to an underestimation and lack of knowledge around the disease (Melanoma skin cancer, 2022). Whilst some countries, such as Australia, have successfully implemented media and awareness campaigns to combat the high incidence rates of ME, most others do not place the same effort in spreading similar knowledge (Rat et al., 2018). As a result, patients often ignore or do not pay enough attention to new spots or changing moles on their body, which can lead to the development of undisturbed skin lesions in hard-to-observe areas such as the back or inner thighs (Montella et al., 2002). Even when a patient detects an unusual skin lesion, scheduling an appointment with a specialist is somehow difficult. A General Practitioner often does not have sufficient skills and experience to diagnose ME, whilst dermatologists have long waiting lists (Tsang & Resneck, 2006). Telemedicine, a branch of eHealth, has been adopted in several countries as a way to increase access to healthcare. It provides the option to schedule a video appointment with the dermatologist, but this can still be time-consuming for the practitioner. Alternatively, medical information and a photo of the skin lesion are forwarded to a specialist who can determine if the lesion is malignant, or whether further investigation is needed (Tensen et al., 2016).   
In recent years, significant advancements in eHealth technology have brought promising opportunities to support patients and healthcare. This includes computer vision, a subfield of AI that focuses on developing models to enable computers to perceive and interpret visual information like humans do (Udrea et al., 2019). EHealth technology has greatly supported the medical sector by providing automatic detection, classification, and diagnosis of medical images with high accuracy, often surpassing the performance of experienced clinicians (Brinker et al., 2019). Substantial research and experiments have been conducted to create performant models over time. The most common medical image to classify ME is obtained through dermoscopy. This is a non-invasive tool that captures characteristics of the epidermis not visible to the naked eye. Several databases of dermoscopy images of different skin lesions are publicly available.

## Machine Learning

The rapid growth of volumes of data generated by various sources has created a pressing need for automated systems that can effectively process and extract insights from it. Machine learning (ML), a subfield of AI, has emerged as a powerful tool for automating data analysis and providing predictions or recommendations based on the patterns it identifies in the data.  
ML models have seen a steady growth and development for ME detection in the past years. Before the widespread of ANN and Deep learning (DL), the most common approach of traditional ML models for ME detection has been to follow a standard workflow in order to set the best ground for the success of the model: lesion segmentation, feature segmentation and extraction, classification (KhakAbi et al., 2012).

Lesion segmentation  
This step consists in separating the skin lesion area of interest, from the normal surrounding skin. Most often these are clearly distinguishable by colour or texture, and it is an important step in the analysis of dermoscopy images as it determines and confines the area where relevant clinical features need to be extracted at a later stage. The border of the segmentation can also provide important hints for determining the nature of the skin lesion.   
This step can be challenging as the dermoscopy images can have poor contrast or non-uniform lighting between the skin lesion and normal skin, or the skin lesion itself can be very similar to the normal skin by nature. There is a broad spectrum of variation within skin lesions that also acts as an obstacle to identify and segment them in a straight-forward manner. Artefacts such as hair, gel, ink, colour calibration charts can be a further challenge to perform this step in a satisfactory way (Appendix - Fig.7). A combination of pre-processing techniques such as elimination of variable lighting effects, contrast enhancement, hair removal or localisation of the lesion can help to perform this step successfully.

Feature segmentation and extraction  
Feature segmentation and extraction is the process of identifying characteristics of the skin lesion image that can help distinguish its nature (Appendix - Fig.8). Common feature examples can be location, visible vascular structures, shades of pink, starburst pattern, dots and globules, etc. For each feature a number of methods are tested through various iterations to determine which one guarantees the highest success for identifying the target feature. These methods can be similar to the one’s in the previous step, examples are light correction, contrast enhancement, colour standardisation, etc.

Classification  
Once the previous steps have been performed, the core part of the problem can be tackled: predicting a skin lesion to be benign or malignant. A most common approach is to follow the ABCD rule of dermatoscopy (Nachbar et al., 1994), based on the criteria of asymmetry (A), border (B), colour (C) and the differential structure (D) which groups the extracted feature within these four areas. Using these attributes in a classifier allows detecting ME with good results. Commonly used classifiers are support vector machines, logistic regression, decision trees, K-nearest neighbours, Bayesian classifiers (Mishra & Celebi, 2016).

In 2009, Alcon et al. (2009) after the pre-processing and segmentation steps, adopted a feature extraction method following the ABCD rule of dermatoscopy (Nachbar et al., 1994). 55 features have been extracted from the skin lesion images to discriminate between benign and malignant lesions, but not all had the same relevance. The best result for the classification step was obtained with a logistic model decision tree. A decision support system tool was created, obtained by combining the outcome of the image classification model with patient-related data known to pose different degrees of risk for ME such as skin type, age, gender, part of the body. An accuracy of 86% with 94% sensitivity and 68% specificity was achieved.  
Karimian et al. (2014) attempted classifying ME from images obtained by conventional digital cameras. They adopted principal component analysis to restrict the feature extraction to 13 relevant ones out of the 137 initially detected, and support vector machines for the classification step. They achieved an accuracy of 82.2%, sensitivity of 77% and specificity of 86.93%.

## Artificial Neural Networks and Deep Learning

The advances in technology have brought ANN to achieve huge popularity in the most recent years. ANN forms the base of DL, a subfield of ML where the algorithms are inspired by the way information is processed by the human brain (Appendix - Fig.9). ANNs take in data and train themselves to recognise the patterns and then predict the output for a new set of similar data. ANNs are composed of layers of neurons which are the core processing units of the network. The initial and final layers are represented by the input layer where the variable enters and the output layer that predicts the outcome. Within a neural network, multiple hidden layers are present, which perform most of the computations required by the network. Neurons of one layer are connected to neurons of the next layer through channels. During the training process of the ANN, each of these channels are assigned a numerical value known as weight. The inputs are multiplied by the corresponding weight and summed to become the input to the neurons in the hidden layer. Each of these neurons are associated with a numerical value called the bias which is then added to the input sum. This value is then passed to a threshold function called the activation function. The result of the activation function determines whether that particular neuron will get activated or not. An activated neuron transmits data to the neurons of the next layer over the channels. By doing so, the data is propagated through the network (forward propagation). In the output layer the neuron with the highest value fires and determines the output. The outcome is basically a probability and the predicted output is compared to the true output to check the accuracy in prediction. During each iteration of the training process, the degree of the error indicates how wrong the current algorithm is, giving an indication of the direction and magnitude of change needed to reduce the error. This information is transferred backward to the network (back propagation) and the weights are adjusted. This cycle of forward and back propagation is iteratively performed with multiple inputs. The process continues until the weights assigned to the network can predict the correct output most of the times (Han et al., 2018).   
In ME classification, one of the greatest benefits of ANN is that it automises the feature segmentation and extraction steps. The model will automatically detect the features that are most relevant for determining the success of the classification through the weights and forward/back propagation process described earlier, making the model more accurate, reliable and less tedious to construct. One of the caveats for benefitting from the use of ANNs is the need for a large amount of data in order to make the model effective.

## Convolutional Neural Networks

Convolutional Neural Networks (CNN) are a specific type of ANN which contains at least one convolutional layer and that specialises in image recognition. This is a particular section of the ANN within the hidden layers that contain filters to perform pattern recognition in the image, such as shape, edges, textures, objects, etc. A filter can be described as a NxN block that slides over the pixels of the image in search of a specified pattern, retaining spatial information, unlike other traditional algorithms. This sliding is actually referred to as convolving, hence the name of the network. It will then give a score for how close the pattern just encountered is to the one it is looking for. The process continues with the filter convolving across each NxN block of pixels from the input until it has covered all of them, and it repeats itself with other filters in search for different patterns. The result will be a collection of numeric arrays that, combined together in a process called pooling, can give a much better understanding of what the image is representing. Subsequent convolutional layers can keep searching for different patterns or features in order to achieve a higher level of understanding. The deeper the network goes, the more sophisticated the search becomes. This allows for superior feature extraction and, generally, an overall better performance (Indolia et al., 2018). A representation of its structure can be visualised in Figure 10 (Appendix - Fig.10).  
Several CNN architectures have been built for image recognition, and some of them reached high popularity due to their impressive results. Competitions in computer vision take place each year to challenge participants to achieve new heights of performance, contributing substantially to the advancements in the field. The dataset on which the challenge is set is most commonly the ImageNet dataset. This consists of “an image database organized according to the WordNet” (an English Lexical Database (WordNet, 2022)) “hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images” (ImageNet, 2022). It is used for research purposes and as a benchmark for image recognition models. It has been shown how a model that performs well on an image recognition task with the ImageNet dataset, is able to transfer its knowledge to other image recognition problems with effective results.  
The most popular CNN architectures can be identified in:

AlexNet   
AlexNet (Krizhevsky et al., 2017) won the 2012 ImageNet ILSVRC-2012 competition (Imagenet Large Scale Visual Recognition Challenge 2022). It is regarded as one of the breakthrough models of the time as it outperformed all competitors by 26.2%, setting a new standard for image recognition performance. Its conception was determined by the availability of larger datasets which prompted the construction of more powerful models and the implementation of more functional techniques to prevent overfitting, such as dropout (Poernomo & Kang, 2018). Because of the complexity and computational demand of the architecture, it was trained simultaneously on two different Nvidia Geforce GTX 580 GPUs, which splits it in two pipelines, resulting in a huge number of parameters: 62.3 million in 5 Convolution layers and 3 fully connected layers.

ResNet  
The winner of ILSVRC-2015 has been ResNet (He et al., 2016), a very deep network with over 100 layers. It introduced the concept of Residual Blocks, making use of “skip connections”. This technique skips some layers by connecting the activations of one layer to another further down the line and forming a so-called residual block. The ResNet is formed by stacking these residual blocks together. The idea behind is that if there is any layer that negatively impacts the performance of the model, it will be skipped by a regularisation process. It allows optimal use of the great amount of layers, improving performance and preventing overfitting/vanishing gradients.

GoogLeNet (Inception)  
Szegedy et al. (Szegedy et al., 2015) built the GoogLeNet architecture, also regarded as Inception, winner of the ILSVRC 2014 competition. This architecture introduced the idea of the inception module to optimise the problem of vanishing gradients by using image distortion, RMSProp (Root Mean Square Propagation) (RMSPROP Optimizer explained, 2021) and batch normalisation techniques. These allow for slower decay of the learning rate which is responsible for preventing convergence of the training and testing models, as well as improve speed, performance, and stability of the model. The Inception architecture is characterised by many small convolutions that lower the number of parameters, bringing them significantly down to 4 million against the 60 million parameters of AlexNet. A classification error of 6.67% has been achieved with a network which is 22 layers deep. Taking in as an input an RGB coloured image of 224x224 pixels, it uses ReLU as an activation function for all intermediate convolutional layers. It finalises its outcome through a softmax layer of one thousand neurons.  
An example of how this model has been implemented for ME classification is given by Yilmaz & Trocan (Yilmaz & Trocan, 2021): they proposed a modified version of the GoogLeNet algorithm to detect malignant ME on the ISIC dataset. They achieved a classification accuracy of 93% and reduced the time complexity by 27 minutes.  
In 2019 Kassani and Kassani (2019) elaborated a comparative study of different deep CNNs for ME classification. After the necessary pre-processing operations on the ISIC 2018 dataset, 4 architectures have been compared: Xception, AlexNet, VGGNet and ResNet. Using transfer learning from the ImageNet pretrained weights for an enhanced performance, the network that achieved the best performance after the pre-processing has been ResNet50, with an accuracy of 92.08 %, a precision of 93.73 %, recall of 92.53 % and F-score of 92.74%. It outperformed the accuracy of the other networks by a value between 2 and 12%, with AlexNet having obtained the lowest result (80.45%).  
Some interesting hybrid models have been proposed that have seen the combination of traditional ML and DL. Haghighi et al. (2020) proposed a method where data augmentation is performed to balance the dataset. It then utilises a CNN architecture of 25 layers, 5 of which are convolutional for the feature extraction step with a support vector machine as a classifier, achieving an accuracy of 89.5%, a sensitivity of 87.7% and a specificity of 91.5%.   
A compelling study has been conducted in 2019 by Brinker et al. (2019) in which the performance of CNN and dermatologist have been compared. The automated system outperformed 136 out of 157 participating dermatologists of all categories of experience in ME detection, achieving a sensitivity of 84.2% and a specificity of 69.2%, versus a result from the clinicians of, respectively, 60 % and 74.1%. It suggests that CV models can positively assist dermatologists in clinical practice for ME detection.

## MobileNet models

The most recent trend has seen a gradual shift from very deep learning models to more computationally efficient ones, where the challenge is to achieve the best balance between performance and computational timing and resources. In this context, several architectures have been developed. MobileNet architectures have captured the attention of researchers as it provides an efficient system with a lightweight design. They represent a class of deep CNN models that are vastly smaller in size and faster in performance than many other popular models. The size difference is due to the number of parameters within these networks and thanks to this, they are characterised by low-latency, low-power required and are considered great DL models to be used on mobile devices. They may lose some points in terms of maximum accuracy compared to the results obtained by the heavy CNN models described earlier, but the trade-off is a substantial gain in computational timing and low amount of resources needed (Howard et al., 2017).

MobileNet v1 introduced depth-wise separable convolutions, which divides the channel and spatial computation in 2 steps. The first one applies a depthwise convolution to perform the filtering stage by applying a convolution to a single input channel at the time, in contrast to standard convolutions that apply convolution to all channels. The second step is called pointwise convolution which performs the linear combination of each of these layers to create the output of the depthwise convolution. The splitting of these operations in 2 steps, makes it lighter than standard convolutions.

MNv2 added an expansion layer in the block to get a system of expansion-filtering- compression using three layers. This has been regarded as an inverted residual block with bottleneck features. The result has been a much quicker model that uses one third of the parameters, has higher accuracy and increased its speed by about 30-40% by computing half of the operations compared to his predecessor (Sinha, 2020).

This architecture has been used also for ME detection. In the study conducted by Wibowo et al. (2020), achieving 70% accuracy by fine-tuning the model on top of the pre-trained architecture. Another interesting example is provided by Indraswari et al. (2022) who reached a remarkable 85% accuracy on the ISIC dataset by adding several layers on top of the convolutional layers of the MobileNetV2. However, sensitivity and specificity scores are suboptimal, as they only achieve a result of 85% each.

Early this year, the third version of MobileNet was released. Some innovations have been introduced, such as a NetAdapt algorithm (Yang et al., 2018) which allows for automated tuning with the device’s CPU through a network architecture search (NAS). Two versions of MobileNet v3 are available: MobileNet v3-Large and MobileNet v3-Small, which are targeted to different availability of resources to use. Although MobileNet v2 had already set a high bar, on ImageNet classification MobileNet v3-Large has proved to be 3.2% more accurate whilst reducing latency by 15%. MobileNet v3-S instead has shown to sacrifice 2.6% in terms of accuracy compared to the MobileNet v2 model, but it reduces latency by 30% (Howard et al., 2019).

In July 2022, during the third International Conference on Applied Engineering and Natural Sciences, Cetiner H. presented a work where MobileNet v3- Large fine-tuned model has been built for skin cancer classification, achieving an impressive accuracy of 98.45% (Çetiner, 2022).

Since early detection of ME is crucial for survival, a system that aids classification and diagnosis of the disease can be extremely useful to improve patient’s outcomes. Mobile phone apps with embedded algorithms that can allow self-screening and detect ME can help towards the scope, therefore increasing survival rates. To the best of our knowledge, a fine-tuned MNv3-S model has not been developed for ME detection. The proposed method aims to utilise transfer learning from MNv3-S and fine-tune it in the attempt of achieving the best performance possible for ME classification.

# Practical Research Methodology

## Definition of requirements

This project intends to develop and compare different ANNs that accurately and consistently detect ME in its early stages, enabling prompt and effective treatment. This requires the elaboration of models that can analyze images of skin lesions and determine whether or not they are indicative of ME.   
Several requirements can be identified for this scope, and they are described in the following sections.

### Data requirements

The success of the model depends on the availability of a comprehensive and diverse dataset for training and validation. The HAM10000 (The HAM10000 dataset, 2018) is a commonly used benchmark dataset for skin lesion classification and widely recognized as a compelling choice thanks to its extensive collection of dermatoscopic images with clinically determined labels. The dataset includes over 10,000 images of size 600x450 pixels, with annotations for 7 different lesion types, of which 1,113 represent ME, allowing for robust evaluation of machine learning models.

### Ethical approval requirements

Typically, research involving human participants requires the study protocol being compliant with ethical conditions. However, this project used data from the HAM10000 dataset, which is a publicly accessible source of non-identifiable medical information that is exempt from the need to seek ethical approval.

### Hardware and software requirements

Building DL models requires access to appropriate hardware and software resources, making it necessary to carefully consider the requirements for both.

Hardware  
In terms of hardware, the availability of a Dell laptop with an Intel Core i7-8550U 1.80GHz -1.99 GHz processor and 8 GB of RAM provides sufficient computing power for the task.

Software

As a self-funded project, it requires to rely on open-source software for efficient modelling tools. There are numerous options available, which can be summarised as follows:

* Python programming language  
  Python is a highly versatile open-source language that has become one of the most widely used programming languages in the world (3.10.4 Documentation, 2022). Created in the 1980s with the goal of making a general-purpose and user-friendly language, Python has since grown in popularity, with a large and supportive community contributing to its development and the creation of many libraries that can be utilized for a range of applications.
* Google Colaboratory  
  Google Colaboratory, or Colab is a cloud-based Jupyter notebook environment that provides a powerful platform for writing and executing Python code. Especially designed for data science and machine learning, it offers free access to GPUs and TPUs, which can significantly accelerate computations for machine learning tasks. (Google Colaboratory, 2022).
* Libraries and frameworks  
  This project requires the use of several Python libraries, which are all open-source and provide essential functionality to build the models. The most relevant are:
  + NumPy is a library for scientific computing that is widely used in the data science and scientific computing communities. It provides efficient and convenient operations for working with arrays, matrices, and linear algebra computations (NumPy Documentation, 2022).
  + Pandas is a powerful library that enables easy manipulation and analysis of data. It provides a flexible and intuitive data structure, the DataFrame, which makes it easy to perform data pre-processing operations and exploratory analysis (Pandas documentation, 2022).
  + TensorFlow is an open-source platform for ML and DL, developed by Google. It provides a comprehensive and flexible ecosystem of tools, libraries, and community resources that allow researchers and developers to build, train, and deploy machine learning models with ease (TensorFlow, 2022).
  + Keras is a high-level neural network library that is written in Python and runs on top of TensorFlow. It is designed to make it easy to build, train, and deploy ANNs, including feedforward and recurrent networks (Keras, 2023).

### Pre-processing requirements

The images in the dataset require pre-processing operations such as resizing or normalisation in order to prepare them for use during the training phase. These steps will be explained in further details in section 3.2.1 - Step 1: Downloading and pre-processing the data.

**3.1.5 Transfer learning requirements**In this project, MNv3-S and MNv2 architectures serve as the foundation for constructing models aimed at detecting ME, capitalising on the knowledge gained from their original assignment to perform the new one more efficiently and with a reduced amount of data. The pre-trained models, in fact, transfer the weights learned from solving their previous problem to be used as a starting point for the current new one. To adapt to the new task, the final layer/s of the pre-trained models are replaced or fine-tuned to achieve better results. This has proven to have a beneficial effect on the performance of the model, as well as saving a considerable amount of time in comparison to creating the model from scratch (Tsiakmaki et al., 2020).

**3.1.6 Model architecture requirements**The foundation of the models is built upon the MobileNet architectures, however, they need to undergo modifications or fine-tuning in order to cater to the specific needs of the ME detection task. This step is to ensure that the models are optimised to perform well on this particular problem and deliver the best results. The modifications include, but are not limited to, adjustments to the number of layers, the regularisation, and the dropout rates, among others.

### Evaluation metrics requirements

All models require to be evaluated using appropriate metrics such as accuracy, sensitivity, specificity in order to assess and compare their performance and ability to detect ME with success. Another metric taken into consideration is the computational timing to run the predictions, as one of the key advantages of lightweight models is their speed.  
Additionally, during step 2 and step 3, the training for each model is recorded (in minutes) to provide a full picture in relation to comparing the computational efficiency of MNv2 and MNv3-S. This metric is discarded in step 4 due to other added computational and fine-tuning operations that would make the timing measurement inconsistent for a comparison to the other models. For a more detailed discussion about the metrics, consult section 3.3 - Collection of results.

## Experiment design

### Step 1: downloading and pre-processing the data

Pre-processing data refers to the cleaning, transforming and organising of data into a format that is suitable for analysis and modelling. It is an important step in the data science pipeline as it helps to ensure that the data is in a consistent and optimal arrangement.   
The HAM10000 dataset can be downloaded from a number of online open-source repositories, such as the ISIC Archive (The HAM10000 dataset, 2018). It comes as three folders, two of which are the images in a standardised JPEG format (divided in part 1 and part 2, respectively of 5,000 and 5,015 instances) along with the corresponding annotations in a separate excel file that include information such as the image name, lesion type, and other metadata.   
The initial pre-processing step involved reorganizing the files to facilitate the training of models. To achieve this, the files were divided into three distinct folders: “train”, “validation” and “test” with each sub-folder labelled with the two relevant classes: “melanoma” or “non-melanoma”.  
This structure enables the model to easily access and distinguish between the different sets of data. At this stage, 8,902 “non-mel” images and 1,113 “mel” images were available, which would make for a notably imbalanced dataset. The optimal number of images required for training a CNN varies based on factors such as model complexity, data quality, and problem type. Considering that transfer learning can provide good performance with smaller datasets, it was decided to include all the ME images and a random sample of the same number of non-ME images, resulting in a balanced dataset with 50% of each class, comprising a total of 2,226 images (Buda et al., 2018).  
The data was divided into three groups: “training”, “validation” and “testing” sets, with a split of 80/10/10%, a commonly used ratio in machine learning. Considering the size of the datasets being large but not massive, this split ratio provides sufficient training data as well as reliable validation to verify over-fitting and fine-tune the models. Finally, by allocating 10% of the data to testing, a sufficient amount of data is reserved to evaluate the generalisation capabilities of the models.

The ImageDataGenerator class is part of Keras library that allows a convenient way to pre-process data in preparation for training a DL model. It provides a high-level interface for configuring the operations required, performing data augmentation and generating batches of data, making it easier to use than implementing the data pre-processing and augmentation steps manually (Tf.keras.preprocessing.image.imagedatagenerator, 2023). The functionalities that have been leveraged for the models include:

Resizing images to a specified target size  
The original MobileNet architecture was trained on images of size 224x224, therefore, to make the most of the pre-trained weights it is recommended to provide it with images of the same size (Howard et al., 2017).  This can be easily achieved by specifying the target size within the ImageDataGenerator() Class.

Batch size  
The batch size for MobileNetV3 can vary depending on the specific use case and the hardware resources available. It determines the number of samples processed by the model in one forward-backward pass and it can affect the training and inference times. A commonly used batch size applied to light-weight pre-trained models such as the MobileNet architectures is 32, therefore it is the one applied (Usmani et al., 2023).

Data augmentation  
Data augmentation is a technique used to increase the size of a dataset by creating new, artificially generated samples by applying random transformations to existing images, such as rotation, scaling, or flipping. This can help to prevent overfitting and improve the generalisation performance of a model (Krizhevsky et al., 2017). This technique has been applied during step 4 when fine-tuning the model.

Normalisation of the pixel values  
The rescale parameter has been used only in step 3 for the model based on MNv2 network. It is, in fact, necessary to feed to the model transformed pixel values from their default range 0-255 to a normalised range between -1 and 1.  
For the MNv3-S based models instead, this step is not necessary because a new feature has been introduced to performs it automatically, the include\_preprocessing  argument in the MNv3-S constructor (Keras, 2023).

### Step 2 - The base model

Given the popularity gained in recent years by the MobileNet architectures, one of the intentions of this project is to compare the latest versions of MNv2 and MNv3-S for ME detection. These two models represent state-of-the-art DL architectures and have been developed with “imageNet” (ImageNet, 2022), a large-scale image dataset used in computer vision research. It contains over 14 million images labelled with more than 20,000 categories, covering a wide range of objects, scenes, and concepts. The dataset is widely used for training and benchmarking computer vision models, particularly deep neural networks, and has played a key role in advancing the field.   
It is often presumed that a newer version of a technological product will overall perform better than its predecessor, but this assumption needs to be verified in several different contexts. Therefore, to answer the first research question of this project - “How does an MNv3-S architecture compare in terms of accuracy, sensitivity, specificity and computational timing compared to MNv2?” - a base model has been created by applying transfer learning on an MNv3-S architecture tailored for ME detection.   
The base structure of the architecture has been forged on top of MNv3-S by taking advantage of transfer learning. The methodology for constructing the base model entailed implementing minimal modifications to the pre-trained MNv3-S architecture in order to preserve its similarity to the original version. To do so, by setting “base\_model.layers[:-1].trainable = False” all layers except the last one have been frozen, which means that their values have been locked from being updated during the training process (He et al., 2016).   
Only two additional layers have been added on top of it, a GlobalAveragePooling2D() layer which reduces the spatial dimensions of the final feature maps from the convolutional layers and shapes it in a correct format for the final Dense layer with 2 neurons that determines the binary output.   
The sigmoid activation function is a popular choice for use in the output layer of a neural network when performing binary classification. It maps the input values to a range between 0 and 1, which represents the predicted probability of each class (Hinton et al., 2006).  
The Adam optimizer is set as the optimisation algorithm. It is a widely used method in DL that derives from the stochastic gradient descent (SGD) and adapts the learning rates of individual parameters based on the historical gradient information. It has several advantages over other optimisation methods. For instance, compared to SGD, Adam is more computationally efficient, less prone to overfitting, and often converges more quickly to a better solution (Barakat & Bianchi, 2021).   
Typically, when using transfer learning, it is a common practice to use a smaller learning rate for the optimizer than when training a model from scratch. This is due to the fact that pre-trained model already contains relevant information about the task, and a smaller learning rate allows the model to learn from the pre-trained weights without the need of significant modifications. Therefore, the learning rate has been set to 0.0001 (Smith, 2017).  
The number of epochs is a critical hyperparameter in DL that determines the number of times the model will see the entire training dataset. For the base model it has been set to 20, which represents a good compromise between the risks of underfitting and overfitting (Prechelt, 2012).

### Step 3 - The comparison model

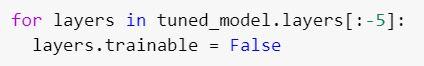
To make the comparison relevant and fair, all parameters have been set in the same way as the base model. In particular, the model maintained:

* The base of MNv2 up until, but excluding, the final layer.
* 2 additional layers on top of the pre-trained model: GlobalAveragePooling2D()  
  Dense, 2 with sigmoid activation function.
* Adam optimiser with a learning rate of 0.0001.
* 'Binary\_crossentropy' loss function.
* 20 epochs.

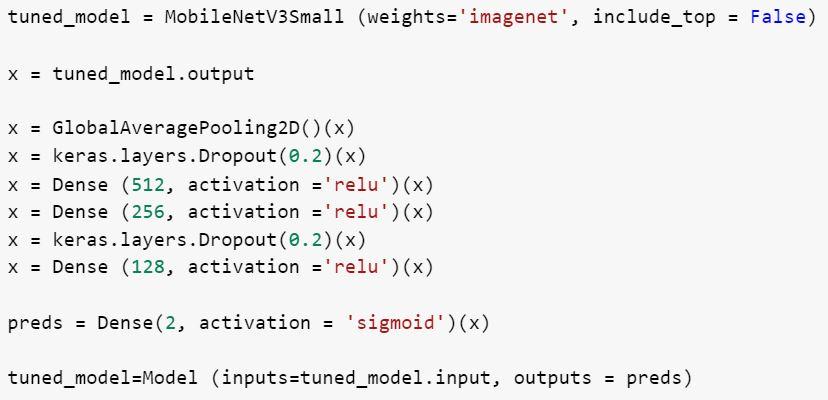
The only difference lies in the need to rescaling the pixel values into a [-1:1] range before feeding the images to the network, an operation that for MNv3-S is performed automatically.

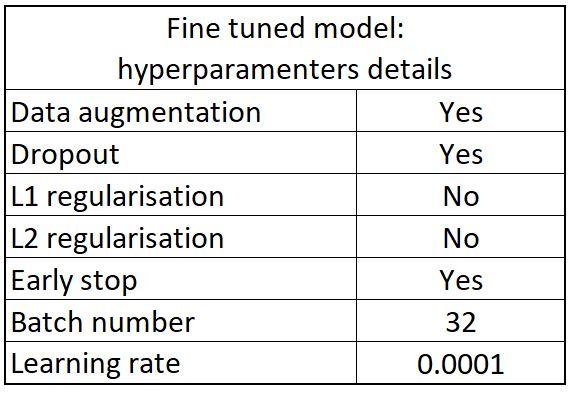
### Step 4 - Fine-tuning the base model

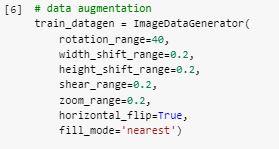
Fine-tuning is the crucial last step to optimise model performance by improving upon the base model. Whilst trying all possible combinations is impractical, multiple iterations and experimentation have been conducted to determine the most effective architecture.   
One of the main parameters to determine for the pre-trained model is the number of layers to freeze. The optimal choice depends on the specific use case, there is no set guideline, but rather a process of trial and error, where multiple options are explored and the performance of each is evaluated on a validation set. The best option in this case has been found to be freezing the top five layers of the base model (Fig 1).

  
Fig 1: code snippet that details the number of frozen layers.

Different hyperparameters have been experimented such as the number of hidden layers (up to a maximum of four to maintain the model light) on top of the base model, the number of neurons in each additional hidden layer (from 128 to 512 in different combinations), and regularisation methods such as data augmentation, batch size, dropout, and L1 and L2 regularisers. Additionally, different values of learning rate have been tested (between 0.001 and 0.00001), and an early call-back set with a patience of 4 has been set to stop the training when the number of epochs would not produce any further improvement.  
The structure of the best performing model can be observed in Figure 2, and the hyper-parameters utilised in Figure 3, among the ones that have been tested. The data augmentation techniques that have been applied can be viewed in Figure 4, and included rotation, width and height shift, zooming, flipping and shear range transformations.

  
Fig. 2: Code snippet that displays the structure of the fine-tuned model.

  
Fig. 3: Hyper-parameters tested during experimentation and eventually included in the fine-tuned model.

  
Fig. 4: Code snippet for applying data augmentation with the ImageDataGenerator class that applies random transformation to the images.

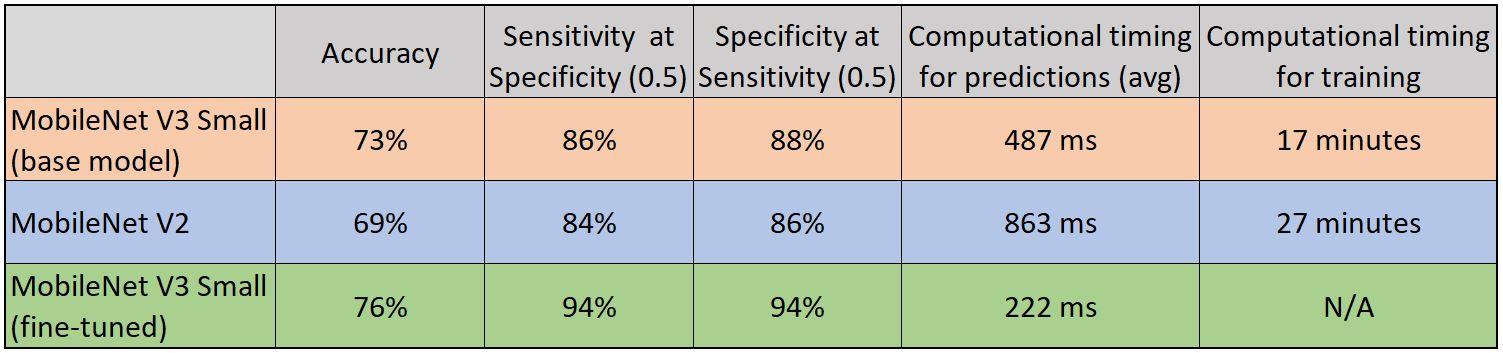
## Collection of results

Accuracy is a common metric used to evaluate the performance of image recognition models. It measures the proportion of correct predictions made by the model over the total number of predictions. For a binary classification problem, accuracy is calculated as the number of true positive and true negative predictions divided by the total number of predictions. It is indeed a good metric for evaluating the performance of image recognition models when the classes are well balanced and have roughly equal number of samples, but it should be complemented with other metrics to get a full picture of the model's performance. In this project the pre-processing steps organised the dataset to be balanced, but in order to have a common ground of comparison with other similar works further metrics have been considered: specificity and sensitivity (Jojoa Acosta et al., 2021).   
The Specificity metric is a measure that evaluates the proportion of correctly predicted negative instances (True Negatives) out of all instances that were actually negative (True Negatives + False Positives). In this specific case it gives an idea of how many images that are not ME are guessed as such.  
The Sensitivity metric, also referred to as Recall or True Positive Rate, assesses the ratio of accurately predicted positive cases (True Positives) relative to the total number of actual positive cases (True Positives + False Negatives). It represents the number of instances that have been guessed as ME when they actually were ME.  
Specificity and sensitivity are closely connected to the ethical issue of over-diagnosing and under-diagnosing ME. This refers to the impact that inaccurate diagnoses can have on patients and society. Over-diagnosing ME refers to the situation where a patient is incorrectly diagnosed with ME, leading to unnecessary anxiety, treatment, and long-term harm to the patient, and is identified by the specificity metric. On the other hand, under-diagnosing ME refers to the scenario where a patient who actually is affected by ME is not diagnosed as such, leading to missed opportunities for treatment and a higher risk of death, identified by the sensitivity metric (Wu and Negbenebor, 2022).   
Recently the DL framework Keras has introduced a "SpecificityAtSensitivity" metric that is used to evaluate the performance of binary classification models. It allows the user to indicate a target sensitivity value and calculates the corresponding specificity that is achieved at that sensitivity. This can be useful for balancing the trade-off between sensitivity and specificity, as these two metrics have an inverse relationship. In the same but opposite way, works the "SensitivityAtSpecificity" metric.   
Generally, calibrating a ME detection model in a more conservatory way would be preferable to identify with the highest rate possible people truly affected by the disease, and, therefore, to limit as much as possible the number of false negatives (patients with ME diagnosed as healthy), as this would pose the highest risk for the life of the patient. On the other hand, setting the bar too high would increase the number of false positives, with possible repercussions that follow a wrong diagnosis. Since the nature of this project is experimental and it aims to compare the performance of different models and architectures, a balanced approach has been followed without favouring either of the two metrics in question (Keras, 2023).

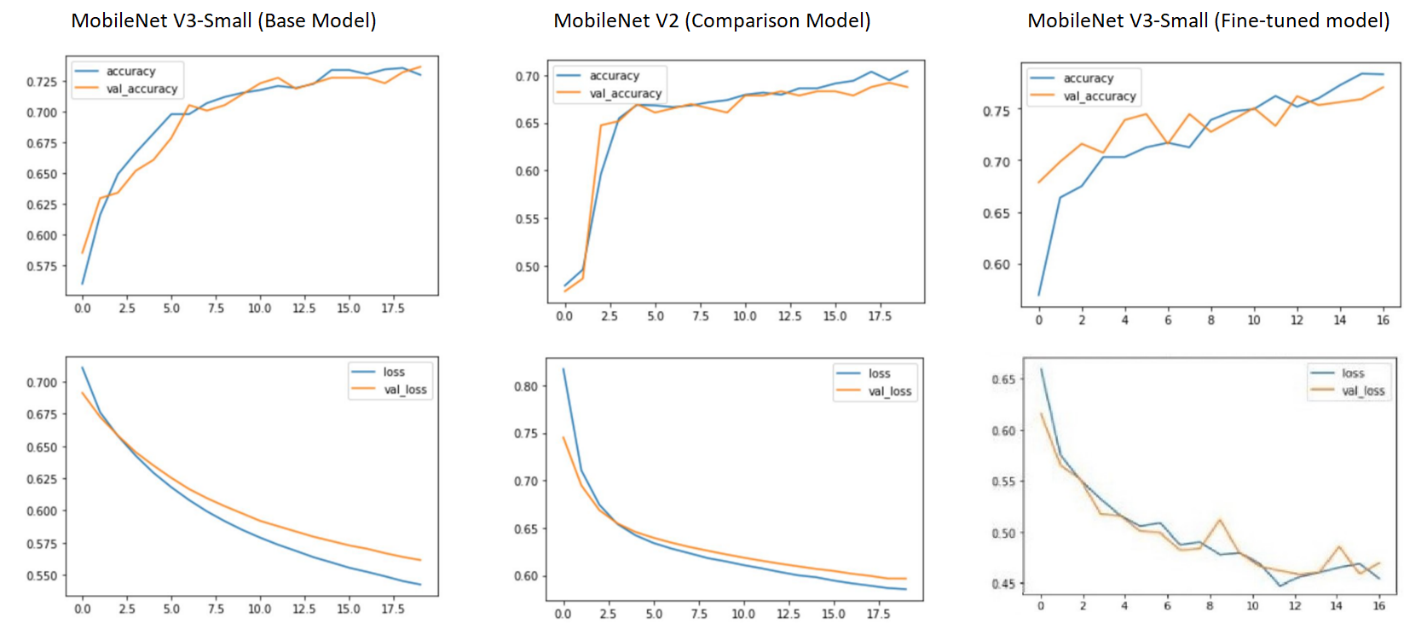
# Results: Analysis and Evaluation

## Analysis of results

The first research question aimed to determine whether the MNv3-S architecture can perform better than the MNv2 architecture in terms of accuracy, sensitivity, specificity and computational timing when applied to a ME classification problem. Steps 2 and 3 of the project development verified this by creating a base model based on MNv3-S and a model with similar features but built upon MNv2.   
The results are presented in Table 1 and demonstrate the performance achieved by the two models.

  
Table 1: An overview of the performance of the models.

It can be clearly observed how MNv3-S outperforms MNv2 in all metrics taken into consideration: it has a substantial edge in the accuracy achieved, 73% against 69%. The sensitivity and specificity scores are closer, with MNv3-S leading by a 2% margin on each of these metrics.   
One major advantage of the MNv3-S model is its faster training time. Whilst the MNv2 required 27 minutes to train, the MNv3-S could be trained in just 17 minutes, a reduction of over 30%. This advantage is further highlighted by the computational timing required for running predictions (Table 1), which is calculated as the average time taken to process the 7 testing batches fed to the models. Notably, the MNv3-S requires only half the time to generate a response compared to the MNv2, making it a highly efficient option.  
By observing the plot of the loss function (Fig.5), it can be noticed how both models converge closely after 20 epochs. The results obtained on the test sets confirm the capability of these models to generalise efficiently on unseen data, as these are very close to the validation test (the confusion matrix and performance on testing sets can be found in the Appendix - Fig.11).

  
Fig. 5: In these plots, the accuracy and loss on training and validation data are represented for each of the models.

In step 4, the fine-tuned model demonstrated superior performance across all metrics compared to the base model. It did in fact, achieve an accuracy of 76% and a sensitivity and specificity of 94% each (Table 1). By incorporating additional hidden layers and implementing regularisation methods, the model trained for 17 epochs before stopping due to the early stop call-back. The loss and accuracy graph demonstrates that the training and validation loss closely converged, indicating a well-performing model capable of generalising effectively (Fig.5). Overall, the fine-tuned model represents a significant improvement over the base model.

## Discussion of results

Based on current research, the classification performance of MNv3-S on the ImageNet dataset has been reported to achieve a Top-1 accuracy of 68.1%, whilst MobileNet V2 achieved a top-1 accuracy of 72% (Sandler et al., 2018). This figure indicates the percentage of test samples for which the correct class is the first prediction made by the model.

The MobileNet models applied in this study for ME classification task during steps two and three yielded results that were comparable to those reported in current literature. Nevertheless, in contrast to previous findings, the experiments here conducted showed that MNv3-S outperformed MNv2 in terms of accuracy. Whilst it's difficult to provide a definitive explanation for this discrepancy, it is possible to speculate that the MNv3-S model may be superior in simple binary tasks, whereas the MNv2 model could achieve higher levels of performance for more complex classifications such as the ImageNet dataset on which it is originally trained upon.

An alternative explanation could be found when observing the plot loss (Fig.5). It can be noted that, although the both converge closely, MNv3-S shows a slightly bigger gap than its counterpart, suggesting a mild degree of overfitting. The predictions run on the test set do not seem to confirm this observation, as they are in line with the results obtained from the validation set, but further testing might be necessary on other data to confirm this hypothesis. In either circumstance, the model was not specifically optimised for the given dataset, as the purpose of the first research question was focused on comparing MNv2 and MNv3-S with fixed parameters rather than achieving optimal performance.   
The key finding of this phase of the project is that MNv3-S can be trained and run predictions significantly faster than V2, with a reduction in time of approximately 30% and 50%, respectively.

Step 4 is a crucial phase in fine-tuning the base model to achieve high levels of performance for ME detection, which addresses the second research question: "What level of accuracy, sensitivity, and specificity can be achieved by fine-tuning the MNv3-S base model?" The results have been promising, with all performance metrics showing improvement (Tab.1). Whilst the accuracy increased from 73% to 76%, a greater gain was noted in sensitivity and specificity, both of which reached 94% with an improvement of 6 to 8%. As highlighted in Section 3.3, these two metrics are vital for addressing over- and under-diagnosis issues in ME detection, making such improvements highly significant (Wu and Negbenebor, 2022). The plot of the training and validation loss function indicates that the model can converge correctly, implying that it is not overfitting (Fig.5).   
Further validation of this model has been obtained through predictions on the testing set, which closely matched the results from the validation set (confusion matrix in Appendix - Fig.11). Moreover, the MNv3-S model also proved to be the fastest in running predictions, cutting the computational time by 50% compared to the base model (Tab.1). These findings make it the best choice among the experiments undertaken and a valuable model to consider for smartphone embedded applications (Fig.6).

Chart, bar chart

Description automatically generated  
Figure 6: A bar chart for an effective comparison of the three models’ performance metrics

This project marks the first time that a fine-tuned MNv3-S architecture has been constructed for ME classification, representing a noteworthy milestone in the ongoing development of DL models, particularly for applications that require efficient use of computational resources.

## Limitations of this study

Whilst the results of this study are promising, there are several limitations to be considered.   
Firstly, the dataset used for training and evaluation could be expanded to provide better generalisation of the model to larger datasets.   
Although the model achieved a good level of accuracy, sensitivity, and specificity, the dataset used in this study may not be representative of the full range of ME conditions and populations, limiting its applicability towards under-represented groups.   
An additional limitation can be identified in the fact that this study only evaluated the performance of the model on one type of device, which may limit the transferability of the findings to other devices.  
Finally, whilst the fine-tuned MNv3-S model proved to be the best choice among the experiments undertaken, other fine-tuned pre-trained models may perform better on this task. Future studies could explore the performance of transfer learning from other pre-trained models and compare their performance with the MNv3-S model.

Despite these limitations, the findings of this study are a significant step forward in developing more accurate and efficient models for ME detection and provide a foundation for further research in this area.

## Future Research

Future research could explore the use of larger datasets to improve the performance of the models. Whilst the current dataset has been found to be adequate, expanding it could provide further insights and improve the model's generalisability. Additionally, the use of transfer learning approaches with other pre-trained architectures could be explored to evaluate their potential for improving the performance of ME detection. Finally, testing the model on a wider range of devices could also provide valuable insights into the feasibility of implementing such models on a larger scale.

# Project Evaluation and Reflection

## Overview of Project and Objectives Review

The objective of this study was to develop a reliable and precise method for identifying ME, which would enable timely and efficient intervention. To achieve this goal, the focus has been placed on creating and comparing various ANNs capable of examining images of skin lesions and ascertaining their correlation with ME. The project was divided into two main research questions. The first research question sought to compare the performance of two different ANNs. Specifically, to determine how an MNv3-S architecture compares to an MNv2 in terms of accuracy, sensitivity, specificity, and computational timing. To answer this question, a well-designed research plan was devised that involved pre-processing, training, and testing both models using a large dataset of skin lesion images, the HAM10000. The results of the experiment showed that the MNv3-S architecture outperforms the MNv2, requiring significant less computational timing.  
The second research question focused on determining the maximum level of performance that could be achieved by fine-tuning the MNv3-S base model. The goal was to determine whether it was possible to further improve the model's accuracy, sensitivity, and specificity by tweaking certain parameters. To answer this question, a series of experiments have been conducted to fine-tune the MNv3-S model and measure its performance. The results showed that the fine-tuned model achieved a significant improvement in accuracy, sensitivity, and specificity compared to the base model, demonstrating the potential for further optimisation.  
Overall, the study objectives were successfully accomplished through a well-designed research plan that was executed in clear and logical steps. The results of the study have important implications for the development of reliable and efficient methods for identifying ME, which could ultimately lead to better outcomes for patients.

## Evaluation of methodology and results

To assess the quality of this study, it is important to consider the thorough approach that has been adopted in planning the experimental design and outlining each step of the project. This has been achieved by conducting a rigorous review of the existing literature, with particular emphasis on the latest research on state-of-the-art models and DL architectures. By utilizing these resources, the study was able to ensure a high level of rigor and precision throughout its execution. The results of this study demonstrate a high level of effectiveness that is both replicable and measured using key metrics such as accuracy, sensitivity, and specificity that have been identified as the most comprehensive and relevant for evaluating this type of study (Jojoa Acosta et al., 2021). They allow for direct comparison with similar studies and provide a valuable contribution to the ongoing effort to improve research in this field. For this reason, the models and experiments developed are both reproducible and publicly available, ensuring that other researchers can build upon the work in a transparent and rigorous manner (Leopaldi, GitHub 2023).   
The use of an established standard method involving the division of data into training, validation, and testing sets helps to prevent bias and ensure the reliability of results. This approach not only helps to minimise overfitting and underfitting, but also allows for more robust evaluation of the model's performance under varying conditions and with unseen data.  
It is important to highlight the relevance of this project for the lives of individuals, as ME is one of the most lethal forms of skin cancer, and early detection is critical for effective treatment and reducing the morbidity and mortality associated with this illness, therefore investing time and resources to improve outcomes for this disease is highly relevant in today's society.

## Alternative aims or hypotheses

In consideration of a wide range of potential research aims and hypotheses to explore, this project could have addressed alternative ideas, for example:

* Using different types of skin lesion datasets to improve performance, robustness and diversity of the model.
* Expanding the scope of the study to include classification of different types of skin lesions beyond ME. This could involve training the network to classify benign lesions, other types of skin cancer, or other dermatological conditions.
* Compare the MNv3-S with the MNv3-Large architecture specifically for ME classification.
* Comparing MNv3-S to other light-weight architectures designed for mobile phones applications, such as EfficientNet or GhostNet.

DL is a rapidly evolving and expansive field, as it continues to develop, there are countless opportunities for researchers to investigate new ideas and advance the understanding of the underlying principles and potential applications of this exciting technology.

## Evaluation of Ethical, Legal, Social, Security and Professional Considerations

The advancement of efficient automated systems for diagnosing ME has significantly aided clinicians in managing this dangerous disease. However, whilst AI can improve the accuracy of ME detection, it cannot guarantee a 100% success rate. Consequently, misdiagnosis remains a possibility and can have grave consequences for patients. As already mentioned in Chapter 3.3, particularly concerning are false-negative cases which can leave a life-threatening ME undetected. Conversely, false positives may result in the unnecessary surgical removal of a benign skin lesion, mistaken for ME. This could lead to a series of complications such as post-surgery issues, aesthetic concerns, financial burden and strain on the healthcare system. From an ethical, social, and professional perspective, all stakeholders involved in the development, deployment, and use of AI-supported technology should aim to minimize the risks of over and under-diagnosis. (Wu and Negbenebor, 2022).  
As AI software is designed to collect, store, and use personal information, addressing privacy issues is both a legal and ethical consideration. Personal health information is strictly private and confidential, and it is protected by rigorous regulations, such as the Data Protection Act (DPA) 2018 in the United Kingdom (Data protection, 2022). Standard practice requires applications to seek user consent, where the risks, limitations, and data handling practices are stated clearly. Often, software applications improve and update their AI algorithms by using new data from users in a de-identified form. The National Statement on Ethical Conduct in Human Research 2007 (Updated 2018) regulates the use of de-identified data, requiring express written consent from the user (National Statement, 2023). However, most patients agree to such contracts without giving proper attention to the terms and conditions and without being aware of what will happen to their personal data. Companies may take advantage of users' negligence and seek unethical use of their personal information, such as selling it to third parties (Jobson, Mar and Freckelton, 2021). Therefore, it is crucial to increase public awareness regarding the handling of personal health information by AI systems and to provide clear and concise information to users. Companies should adhere to ethical standards and legal requirements while respecting users' privacy rights. Additionally, regulators and policymakers should enact strict laws and enforce them rigorously to ensure that the collection, storage, and use of personal data by AI systems are transparent and respectful of individual rights.  
Skin cancer is more prevalent in the Caucasian demographics, and therefore, skin lesion datasets are often gathered by predominantly Caucasian countries. This can lead to a lack of representation of minority groups and ethnicities, resulting in inaccurate outcomes for these populations and contributing to healthcare inequality and disparities. To address this issue, there is a pressing need to create larger datasets that are more inclusive and representative of minorities. This may require significant effort and resources; however, it will lead to more accurate and equitable results and ultimately benefit all patients, regardless of their ethnicity. In the long run, it is hoped that the issue of underrepresentation can be limited by the improved and more inclusive datasets that will result from such efforts (Wu and Negbenebor, 2022).

## Personal reflection

Working on this ME detection study has been both challenging and rewarding. As someone who is passionate about using technology to improve healthcare, it was exciting to conduct a project that aimed to develop a DL model that could assist in early detection of ME. However, it was not without obstacles. The project required a lot of research and experimentation to identify the most effective algorithm and features for the model. Balancing my work, studies, and family commitments whilst working under tight deadlines was a challenge, but with a well-organized plan in place, I was able to manage my time effectively and complete the project successfully.  
Despite the difficulties, the results were highly satisfying. The final model achieved a high level of accuracy, sensitivity, and specificity, which indicates that it can identify ME with a high degree of reliability. This was particularly rewarding because it meant that this work could potentially make a positive impact on people's lives by facilitating earlier and more accurate diagnoses of ME, a disease that can be life-threatening if not caught early.  
This project was an opportunity for me to apply my technical skills in an area that I am passionate about and to work towards improving people's health outcomes. It marks the final step towards completing my MSc degree, but I recognise that it is only the beginning of my journey as a Data Scientist. I look forward to further honing my skills and knowledge, and to applying them to future projects that can make a meaningful impact on people's lives.

# Conclusion and Recommendations

This study aimed to establish a reliable and accurate approach to identify ME, facilitating prompt and effective intervention. To attain this objective, the study focused on constructing and comparing multiple ANNs based on the Mobile Net networks capable of analyzing images of skin lesions and determining their association with ME. The research was structured around two primary questions. The first sought to determine how an MNv3-S architecture compares to an MNv2 in terms of accuracy, sensitivity, specificity, and computational timing, and the second focused on determining the maximum level of performance that could be achieved by fine-tuning the MNv3-S base model.  
These have been effectively addressed and answered through a comprehensive research design that incorporated a sequence of experiments carried out in a clear and systematic manner, with each step corresponding to a crucial phase of the process:

* Step1 - Data acquisition and pre-processing.
* Step2 - Creation of the base model.
* Step 3 - Creation of the comparison model.
* Step 4 - Fine-tuning of the final model.

The analysis demonstrated that the model based on MNv3-S, which was modified to the minimum extent required to adapt to a binary ME detection task (base model), outperformed the comparison model constructed in the same way but using MNv2. The obtained accuracy of 73% is 4% higher than the comparison model, whilst sensitivity and specificity are respectively 86% and 88%, an improvement of 2% each. Notably, the MNv3-S based model also exhibited significantly faster training and prediction times, with reductions of 30% and 50%, respectively.  
The addition of four dense hidden layers to the network, along with data augmentation and regularisation techniques such as dropout and early stopping, resulted in further improvement of the final fine-tuned model. This enhanced the model's accuracy to 76% and its specificity and sensitivity to 94%. Additionally, the time required for predictions was halved.  
In conclusion, MNv3-S outperforms its predecessor MNv2 for ME binary classification by offering better performance and reduced computational timing. By designing a further improved fine-tuned model, MNv3-S can deliver an overall satisfactory performance, marking a significant advancement in exploring the potential of this architecture for ME detection. This study demonstrates that MNv3-S is a valuable and efficient architecture for creating lightweight and high-performing models.  
A study such as this one must consider ethical, legal, social, and professional implications. It is crucial to minimise the risk of under or over-diagnosis when developing ME prediction models. The handling of personal information must comply with the current data and privacy regulations. Furthermore, the representation of minority groups and ethnicities in datasets should be taken into account to avoid inaccurate outcomes and contribute to healthcare inequality and disparities (Wu and Negbenebor, 2022).  
Although these considerations could pose limitations to this study, as well as the fact that it is built on a relatively small dataset and the evaluation of the model's performance was ran on only one device, it does offer encouraging outcomes that can direct further investigation in this domain.  
Future research can focus mainly on three key areas to enhance the performance and generalisability of the ME detection model: using larger datasets, exploring transfer learning approaches with other pre-trained models, and testing the model on a wider range of devices for scalability.

# Reference List

* Barakat, A. and Bianchi, P. (2021) “Convergence and dynamical behavior of the Adam Algorithm for nonconvex stochastic optimization,” *SIAM Journal on Optimization*, 31(1), pp. 244–274. Available at: https://doi.org/10.1137/19m1263443.
* Brinker, T., Hekler, A., Enk, A., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., von Kalle, C., Fröhling, S., Schilling, B. and Utikal, J., 2019. Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119, pp.11-17.
* Buda, M., Maki, A. and Mazurowski, M.A. (2018) “A systematic study of the class imbalance problem in Convolutional Neural Networks,” *Neural Networks*, 106, pp. 249–259. Available at: https://doi.org/10.1016/j.neunet.2018.07.011.
* Çetiner, H. (2022) *MobileNetV2 ve Mobilenetv3 Tabanlı Derin öğrenme Yaklaşımları Ile Cilt kanserlerinin sınıflandırılması*, *EasyChair Home Page*. EasyChair. Available at: https://www.easychair.org/publications/preprint/zbTM5 (Accessed: : February 26, 2023).
* Colab.research.google.com. 2022. *Google Colaboratory*. [online] Available at: <https://colab.research.google.com/> [Accessed : February 26, 2023].
* Docs.python.org. 2022. *3.10.4 Documentation*. [online] Available at: <https://docs.python.org/3/> [Accessed : February 26, 2023].
* Fernandez Alcon, J., Ciuhu, C., ten Kate, W., Heinrich, A., Uzunbajakava, N., Krekels, G., Siem, D. and de Haan, G., 2009. Automatic Imaging System With Decision Support for Inspection of Pigmented Skin Lesions and Melanoma Diagnosis. *IEEE Journal of Selected Topics in Signal Processing*, 3(1), pp.14-25.
* Leopaldi G., GitHub 2023. *GitHub*. Available at: https://github.com/giulioleop/Melanoma-Detection-with-MobileNet-V3 (Accessed: February 26, 2023).
* GOV.UK. 2022. *Data protection*. [online] Available at: <https://www.gov.uk/data-protection#:~:text=The%20Data%20Protection%20Act%202018%20is%20the%20UK's%20implementation%20of,used%20fairly%2C%20lawfully%20and%20transparently> [Accessed : February 26, 2023].
* Haghighi, S., Danyali, H., Helfroush, M. and Karami, M., 2020. A Deep Convolutional Neural Network for Melanoma Recognition in Dermoscopy Images. *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*.
* Han, S.-H. *et al.* (2018) “Artificial Neural Network: Understanding the basic concepts without mathematics,” *Dementia and Neurocognitive Disorders*, 17(3), p. 83. Available at: https://doi.org/10.12779/dnd.2018.17.3.83.
* Harvard Dataverse. 2018. *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*. [online] Available at: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T> [Accessed : February 26, 2023].
* He, K. *et al.* (2016) “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. Available at: https://doi.org/10.1109/cvpr.2016.90.
* Hinton, G.E., Osindero, S. and Teh, Y.-W. (2006) “A fast learning algorithm for deep belief nets,” *Neural Computation*, 18(7), pp. 1527–1554. Available at: https://doi.org/10.1162/neco.2006.18.7.1527.
* Hosseinzadeh Kassani, S. and Hosseinzadeh Kassani, P., 2019. A comparative study of deep learning architectures on melanoma detection. *Tissue and Cell*, 58, pp.76-83.
* Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.
* Howard, A. *et al.* (2019) “Searching for MobileNetV3,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* [Preprint]. Available at: https://doi.org/10.1109/iccv.2019.00140.
* *ImageNet (2022)*. Available at: https://image-net.org/index (Accessed: February 26, 2023).
* *Imagenet Large Scale Visual Recognition Challenge (ILSVRC)* (no date) *ImageNet*. Available at: https://image-net.org/challenges/LSVRC/ (Accessed: : February 26, 2023).
* Indolia, S. *et al.* (2018) “Conceptual understanding of convolutional neural network- a deep learning approach,” *Procedia Computer Science*, 132, pp. 679–688. Available at: https://doi.org/10.1016/j.procs.2018.05.069.
* Indraswari, R., Rokhana, R. and Herulambang, W., 2022. Melanoma image classification based on MobileNetV2 network. *Procedia Computer Science*, 197, pp.198-207.
* Jobson, D., Mar, V. and Freckelton, I., 2021. Legal and ethical considerations of artificial intelligence in skin cancer diagnosis. *Australasian Journal of Dermatology*, 63(1).
* Jojoa Acosta, M.F. *et al.* (2021) “Melanoma diagnosis using Deep Learning techniques on dermatoscopic images,” *BMC Medical Imaging*, 21(1). Available at: https://doi.org/10.1186/s12880-020-00534-8.
* Karimian, A., Ramezani, M. and Moallem, P., 2014. Automatic Detection of Malignant Melanoma using Macroscopic Images. *Journal of Medical Signals &amp; Sensors*, 4(4), p.281.
* Keras.io, (2022). *Keras Documentation: Keras API reference*, *Keras*. Available at: https://keras.io/api/ (Accessed: February 26, 2023).
* Keras.io, (2022). *Keras Documentation: Mobilenet, mobilenetv2, and MobileNetV3*, *Keras*. Available at: https://keras.io/api/applications/mobilenet/ (Accessed: February 26, 2023).
* KhakAbi, S. *et al.* (2012) “Multilevel feature extraction for skin lesion segmentation in dermoscopic images,” *SPIE Proceedings* [Preprint]. Available at: https://doi.org/10.1117/12.911664.
* Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) “ImageNet classification with deep convolutional Neural Networks,” *Communications of the ACM*, 60(6), pp. 84–90. Available at: https://doi.org/10.1145/3065386.
* Leopaldi G. (2022) “Transfer learning and fine-tuning of MobNet V3 architecture for Melanoma detection”, PROM 04.
* Mishra, N.K. and Celebi, E. (2016) “An overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning.”
* Montella, M. *et al.* (2002) “An assessment of factors related to tumor thickness and delay in diagnosis of melanoma in Southern Italy,” *Preventive Medicine*, 35(3), pp. 271–277. Available at: https://doi.org/10.1006/pmed.2002.1067.
* Nachbar, F., Stolz, W., Merkle, T., Cognetta, A., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O. and Plewig, G., 1994. The ABCD rule of dermatoscopy. *Journal of the American Academy of Dermatology*, 30(4), pp.551-559.
* National Statement, 2023. [online] Available at: <https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-condhttps://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018uct-human-research-2007-updated-2018> [Accessed : February 26, 2023].
* nhs.uk. 2022. *Melanoma skin cancer*. [online] Available at: <https://www.nhs.uk/conditions/melanoma-skin-cancer/> [Accessed : February 26, 2023].
* Numpy.org. 2022. *NumPy Documentation*. [online] Available at: <https://numpy.org/doc/> [Accessed : February 26, 2023].
* Pandas.pydata.org. 2022. *Pandas documentation — pandas 1.4.2 documentation*. [online] Available at: <https://pandas.pydata.org/pandas-docs/stable/> [Accessed : February 26, 2023].
* Poernomo, A. and Kang, D.-K. (2018) “Biased dropout and Crossmap dropout: Learning towards effective dropout regularization in Convolutional Neural Network,” *Neural Networks*, 104, pp. 60–67. Available at: https://doi.org/10.1016/j.neunet.2018.03.016.
* Prechelt, L. (2012) “Early stopping — but when?,” *Lecture Notes in Computer Science*, pp. 53–67. Available at: https://doi.org/10.1007/978-3-642-35289-8\_5.
* Rat, C. *et al.* (2018) “Use of smartphones for early detection of melanoma: Systematic review,” *Journal of Medical Internet Research*, 20(4). Available at: https://doi.org/10.2196/jmir.9392.
* *RMSPROP Optimizer explained* (2021), *Machine learning journey*. Available at: https://machinelearningjourney.com/index.php/2021/01/06/rmsprop/ (Accessed : February 26, 2023).
* Sandler, M. *et al.* (2018) “MobileNetV2: Inverted residuals and linear bottlenecks,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* [Preprint]. Available at: https://doi.org/10.1109/cvpr.2018.00474.
* Sarker, I.H. (2021) “Deep cybersecurity: A comprehensive overview from neural network and Deep Learning Perspective,” *SN Computer Science*, 2(3). Available at: https://doi.org/10.1007/s42979-021-00535-6.
* Sinha, S. (2020) *Why google's mobilenetv2 is a revolutionary next gen on-device computer vision network*, *Analytics India Magazine*. Available at: https://analyticsindiamag.com/why-googles-mobilenetv2-is-a-revolutionary-next-gen-on-device-computer-vision-network/#:~:text=How%20Is%20It%20Different%20From,on%20a%20Google%20pixel%20phone. (Accessed : February 26, 2023).
* Smith, L.N. (2017) “Cyclical learning rates for training neural networks,” *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* [Preprint]. Available at: https://doi.org/10.1109/wacv.2017.58.
* Szegedy, C. *et al.* (2015) “Going deeper with convolutions,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. Available at: https://doi.org/10.1109/cvpr.2015.7298594.
* TensorFlow. 2022. *TensorFlow*. [online] Available at: <https://www.tensorflow.org/> [Accessed : February 26, 2023].
* *Tf.keras.preprocessing.image.imagedatagenerator  :   tensorflow V2.11.0* (no date) *TensorFlow*. Available at: https://www.tensorflow.org/api\_docs/python/tf/keras/preprocessing/image/ImageDataGenerator (Accessed: February 26, 2023).
* The International Agency for Research on Cancer (IARC) (no date) *Global cancer observatory*, *Global Cancer Observatory*. Available at: https://gco.iarc.fr/ (Accessed : February 26, 2023).
* Tensen, E. *et al.* (2016) “Two decades of teledermatology: Current status and integration in National Healthcare Systems,” *Current Dermatology Reports*, 5(2), pp. 96–104. Available at: https://doi.org/10.1007/s13671-016-0136-7.
* Tsang, M.W. and Resneck, J.S. (2006) “Even patients with changing moles face long dermatology appointment wait-times: A study of simulated patient calls to dermatologists,” *Journal of the American Academy of Dermatology*, 55(1), pp. 54–58. Available at: https://doi.org/10.1016/j.jaad.2006.04.001.
* Tsiakmaki, M. *et al.* (2020) “Transfer learning from Deep Neural Networks for predicting student performance,” *Applied Sciences*, 10(6), p. 2145. Available at: https://doi.org/10.3390/app10062145.
* Udrea, A., Mitra, G., Costea, D., Noels, E., Wakkee, M., Siegel, D., Carvalho, T. and Nijsten, T., 2019. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *Journal of the European Academy of Dermatology and Venereology*, 34(3), pp.648-655.
* Usmani, I.A. *et al.* (2023) “Interactive effect of learning rate and batch size to implement transfer learning for Brain tumor classification,” *Electronics*, 12(4), p. 964. Available at: https://doi.org/10.3390/electronics12040964.
* Wibowo, A., Adhi Hartanto, C. and Wisnu Wirawan, P. (2020) “Android Skin Cancer Detection and classification based on MobileNet V2 model,” *International Journal of Advances in Intelligent Informatics*, 6(2), p. 135. Available at: https://doi.org/10.26555/ijain.v6i2.492.
* *WordNet* (2022) *Princeton University*. The Trustees of Princeton University. Available at: https://wordnet.princeton.edu/ (Accessed: February 26, 2023).
* Wu, J. and Negbenebor, N., 2022. Melanoma Screening: The Ethics of Over- and Underdiagnosis. *Rhode Island Medical Journal*, (April issue).
* Yang, T.-J. *et al.* (2018) “NetAdapt: Platform-aware neural network adaptation for mobile applications,” *Computer Vision – ECCV 2018*, pp. 289–304. Available at: https://doi.org/10.1007/978-3-030-01249-6\_18.
* Yilmaz, E. and Trocan, M. (2021) “A modified version of googlenet for melanoma diagnosis,” *Journal of Information and Telecommunication*, 5(3), pp. 395–405. Available at: https://doi.org/10.1080/24751839.2021.1893495.

1. **Research Proposal**The research proposal, which was developed as the final assignment of PROM04, can be found on [Github](https://github.com/giulioleop/PROM04-Research-proposal.git) at this link:  
   <https://github.com/giulioleop/PROM04-Research-proposal.git>
2. **Code for this project**The code for this project is included on [Github](https://github.com/giulioleop/Melanoma-Detection-with-MobileNet-V3) at this link:  
   https://github.com/giulioleop/Melanoma-Detection-with-MobileNet-V3
3. **Figure 7**A picture containing application

   Description automatically generatedFig.7: Different artifacts in Dermoscopy images (Mishra & Celebi, 2016)
4. **Figure 8  
     
   A collage of different foods

   Description automatically generated with low confidence** Fig.8: Dermoscopy images illustrating a variety of skin lesions (Mishra & Celebi, 2016)
5. **Figure 9**  
     
   **Diagram

   Description automatically generated**  
   Fig.9: Architecture of a Neural Network
6. **Figure 10**  
     
   **Diagram

   Description automatically generated**  
   Fig.10: Representation of a CNN (Sarker, 2021)
7. **Figure 11  
     
   Table

   Description automatically generated**Fig.11: Test set predictions: computational timing, performance and confusion matrix from all three models.