Recognition of Nevi on Human Body in Internet Images

by
Dimitrios Syrigos
A thesis submitted in partial fulfillment of the
requirements for the degree of

Electrical and Computer Engineering

Technical University of Crete

2021

Examining Committee in Charge:

1. Professor Michael Zervakis (Supervisor)
2. Professor Costas Balas
3. Professor Euripides Petrakis

Declaration

I, Dimitrios Syrigos, hereby declare that the work presented in this thesis is completely my own and has not been submitted for any other degree qualification. Parts of the work have been previously published in conference or journal papers, this has been mentioned in the thesis.Where I have consulted the work of others, this is always clearly stated.

Contents

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis supervisor Pr.Michael Zervakis of the Department of Telecommunications and deputy dean at Technical University of Crete for his patience and his continuous advice and encouragement throughout this thesis. Also I would like to thank Mr.Alexandros Christodoulakis and Mr.Harris Karanikas from Datamed S.A for the original idea.
Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

ABSTRACT

Skin cancer is one of the deadliest forms of cancer. After it metastasizes from its origin into other tissues, the response rate to
treatment declines as low as 5%, and its 10-year survival rate is only about 10%. After metastasis, there is no surgical removal option available for treatment.However, an early diagnosis and a surgery removal, significantly increase the probability of survival. Dermoscopy is a noninvasive high-resolution imaging technique that assists physicians in making more accurate diagnoses of skin cancers.
Therefore, this thesis proposes highly accurate methods, from three different approaches, regarding the skin lesion segmentation(i.e., isolating the lesion from the rest of the image) and classification of nevus and malignant skin lesions. The main point is to build a system that will be able to identify potentially dangerous cases.
We explored through relevant datasets, the effectiveness of both pre-trained and scratch built models with and without segmented images, where the skin lesion area has been isolated as well as with and without cross validation methods.
In the end, the results obtained from all these classifiers and approaches are also compared.
The study showed that the implementation of Deep learning within the field of cancer diseases can be the most suitable way to classify and recognize skin cancer images, which can be very beneficial in the field of medicine for early diagnosis and improve the accurate diagnosis result. This current work showed an output result of 91% accuracy.

Chapter 1
Introduction

1.1 Motivation

Skin diseases are categorized into many different classes and subclasses. Melanoma is a potentially fatal type of skin cancer which is often misdiagnosed as benign or left undiagnosed. In 2018, it was estimated that there would be 91,270 new reported cases of melanoma and approximately 9,320 lost lives in the United States (Siegel et al., 2018). One in 27 men and one in 42 women were predicted to develop melanoma in their lifetimes (Siegel et al., 2018), up from 1:33 and 1:52, respectively, in 2016 (Spiegel et al., 2016). The direct economic cost of melanoma treatment is reported to be over $3.3 million dollars in the US annually (Guy et al., 2015). The indirect cost of melanoma (premature mortality) is much higher and is estimated to be over $3 billion (Guy Jr & Ekwueme, 2011). Early detection, in addition to saving lives, helps to reduce these costs. Despite the unclear perspective, a melanoma diagnosis is not necessarily a fatal diagnosis. Early detection can increase life expectancy (Freedberg et al., 1999). Patients who receive an early diagnosis have a 98% 5Year relative survival rate, whereas in case a patient is diagnosed in later stages, has only a 17% survival rate (Spiegel et al., 2016). Early detection is vital; the lives of patients are at

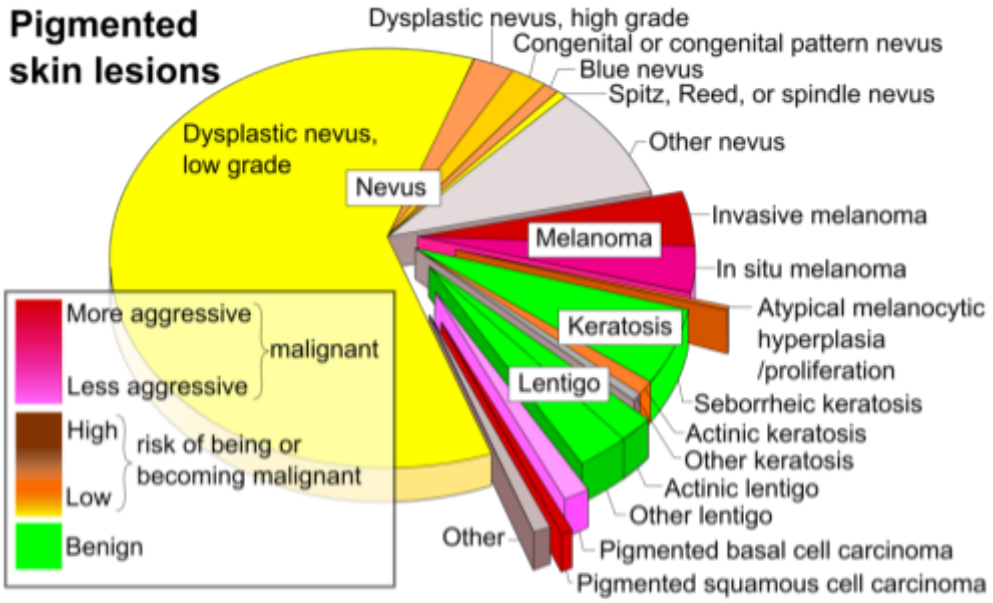


Figure 1.1 Skin Lesion Taxonomy
Source:https://en.wikipedia.org/wiki/Congenital_melanocytic_nevus

the mercy of an accurate diagnosis.Dermatologists have developed guidelines to improve the diagnosis of skin lesions, such as: the ABCD rule (Nachbar et al., 1994), the 7-point checklist (Argenziano et al., 1998), and the CASH algorithm (Henning et al., 2007). Though these methodologies may help with diagnosis, they are imperfect and subjective. Dermatologists often rely on personal experience and evaluate lesions on a case-by-case basis.

When inspecting a lesion, a dermatologist will often take into account the patient's local lesion pattern in comparison to the entire body (Gachon et al., 2005). As a result, without any computer-based assistance, the clinical diagnosis true positive rate for melanoma detection is reported to be around 65-80% (Argenziano et al., 1998). If dermoscopic images are used and the professional has received formal education (Binder et al., 1997), the diagnosis of skin lesions may be improved by 5-30% (Garnavi et al., 2011).

However, the visual differences between benign and melanoma skin lesions can be extremely subtle (Figure 2.5) and their differentiation can be exceptionally difficult, even for trained professionals, thus the success of these methods is limited [(Argenziano & Soyer, 2001), (Kittler et al., 2002), (Vestergaard et al., 2008)]. Due to the severity of melanoma, the significance of early diagnosis and the shortage of trained professionals in some regions [(Brown, 2015),(Australian Medical Association et al, 2005)], as well as the less than perfect unhelpful classification methods, there is a strong motivation to develop and utilize computer aided diagnosis systems (CADx) to aid in the classification of skin lesions.

In this project, we are interested in investigating systems like that, specifically, an intelligent medical imaging-based skin lesion diagnosis system which would help determine whether a dermoscopic image of a skin lesion contains a nevus (benign) or a non nevus skin lesion.

1.2 Problem Statement

The fundamental problem addressed by this thesis can be stated as a question : How do we use the latest developments in deep learning in order to implement a two-class classifier, that is capable of examining an image ,containing a skin lesion and predicting an outcome (nevus or non nevus) with a high enough degree of confidence so as to enhance current early malignant detection methods ?

More specifically, it is desirable to have an intelligent model that indicates how much nevus skin lesions differ from nevus models. Such a model must predict - based on RGB images of skin moles- the occurrence of nevus and other types of skin lesions (non nevus), which are needed to be examined further from a doctor. This problem results from a company's (DataMed) need to create a product, which would be capable of downloading skin images from the web and classifying them to the aforementioned categories.

The skin images that will be considered as non-nevus lesions will be processed by another deep learning model, which is a melanoma detector. Our work is to develop an accurate nevus/non-nevus classifier with a target to minimize the misclassification errors. In order to complete our work, we create a nevus/non-nevus classifier by using pretrained networks for classification, which we feed them with datasets suitable for this purpose.

After this task, we create a skin lesion segmentation system, which isolates the lesion surface on the skin,in this way we create a new dataset, which is like the first ones we used but with segmented images through the skin lesion segmentation system.

In the end,comparing our approaches helps us see which of them provides us with the most accurate results.

The DataMed's plan of the product is illustrated in the following image.

We train our neural networks with images from a relevant database in order to make a reliable diagnosis between nevus and non nevus, so that later this result can receive further evaluation by the company's algorithms as to whether it is melanoma or not.

1.3 Main Contributions

The main contributions of this thesis include:

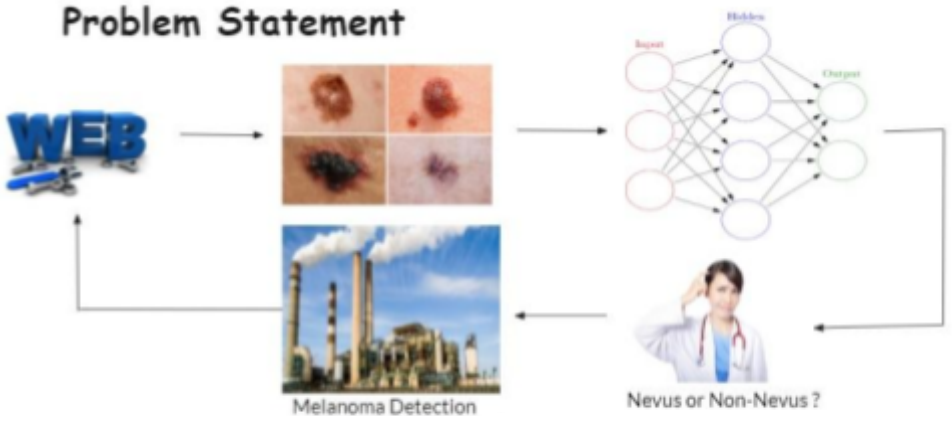


Figure 1.2 Problem Statement

- Design, implementation, and evaluation of a CNN fully functional solution for nevus/non nevus classification.
- Implementation and evaluation of multiple approaches to nevus/non nevus classification using transfer learning models.
- Implementation and evaluation of a skin lesion segmentation system.
- A comparison of methods with and without segmented datasets and discussion of the results.

The work in this thesis indicates that the problem of nevus detection, segmentation and classification, can be facilitated from large datasets with multiple examples.

1.4 Background

Artificial intelligence (AI) is a broad term, encapsulating many technologies, that is steadily being integrated into industrial settings. The majority share of AI funding is currently being invested in the health care sector (see figure 2.1), where 1.5 Billion dollars have been invested in nearly 190 healthcare focused AI startups in the previous five years (*Ai in healthcare heatmap: From diagnostics to drug discovery startups, the category heats up.*, n.d.). One third of the healthcare startups receiving funding after January 2015 were working in imaging and diagnostics (*Ai in healthcare heatmap: From diagnostics to drug discovery startups, the category heats up.*, n.d.). There are many promising, and sometimes surprising use cases for AI in medical imaging and diagnostics.

Some recent advancements include: retinal images to predict cardiovascular risk factors (Poplin et al., n.d.), identifying diabetic retinopathy from retinal fundus photographs (Gulshan et al., 2016), identifying tumors from pathology slides (Liu et al., 2017), and detecting tuberculosis on chest radiographs (Lakhami & Sundaram, 2017). Despite the promise and excitement of AI in healthcare and deep learning in general (see Garnter hype cycle in Figure 2.2), AI does not always bring positive sentiment. There are speculations and concerns about the integration of AI into the medical industry, and to what extent AI should be used (Mukherjee., 2017).

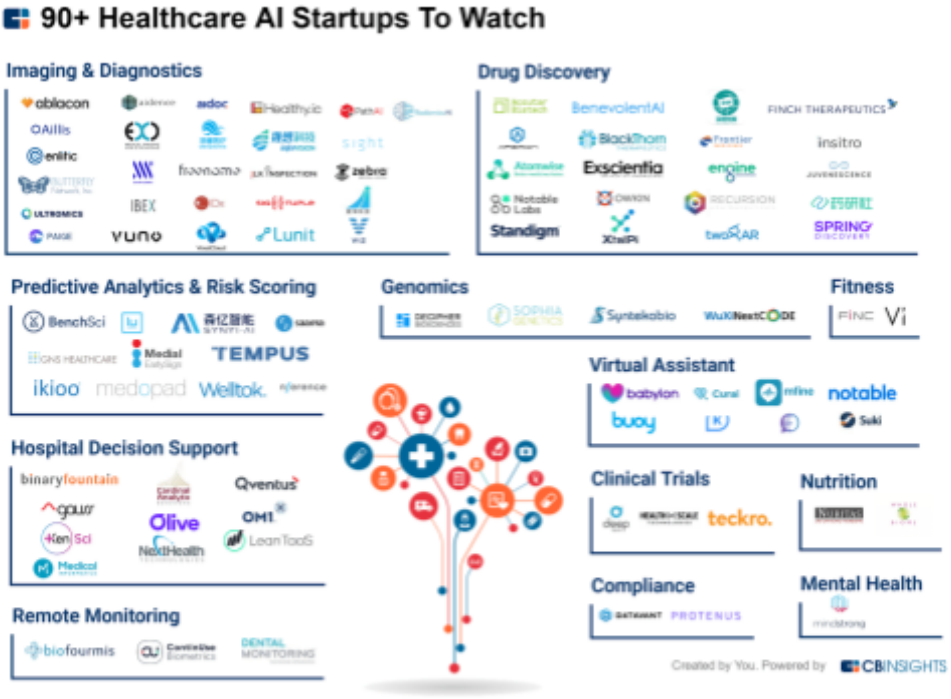


Figure 1.3:Map of start up companies in the AI+Healthcare field
Source:https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/

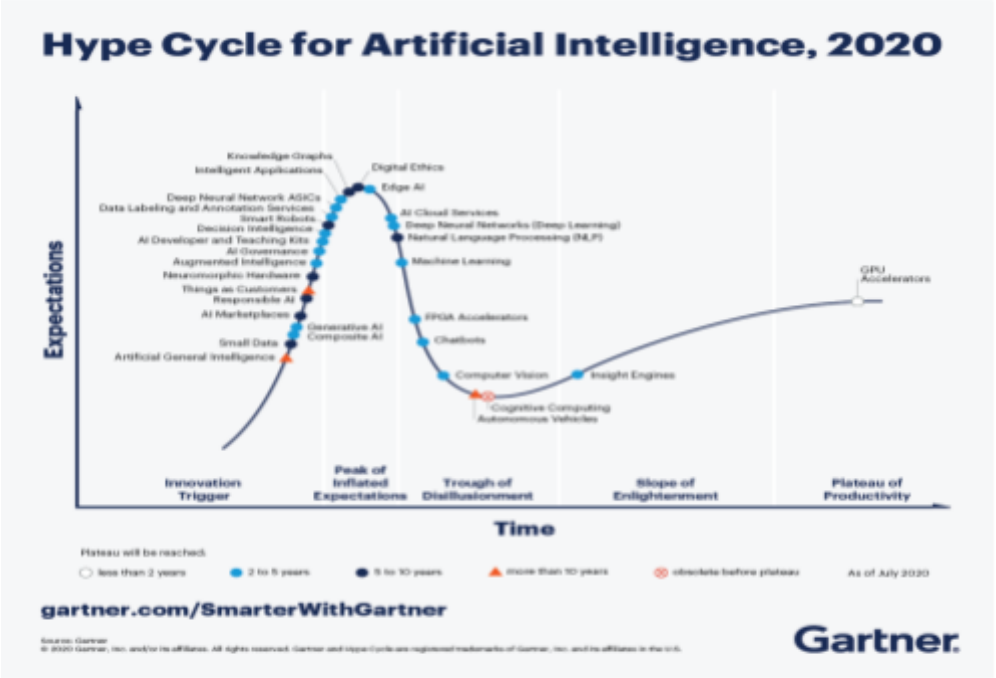


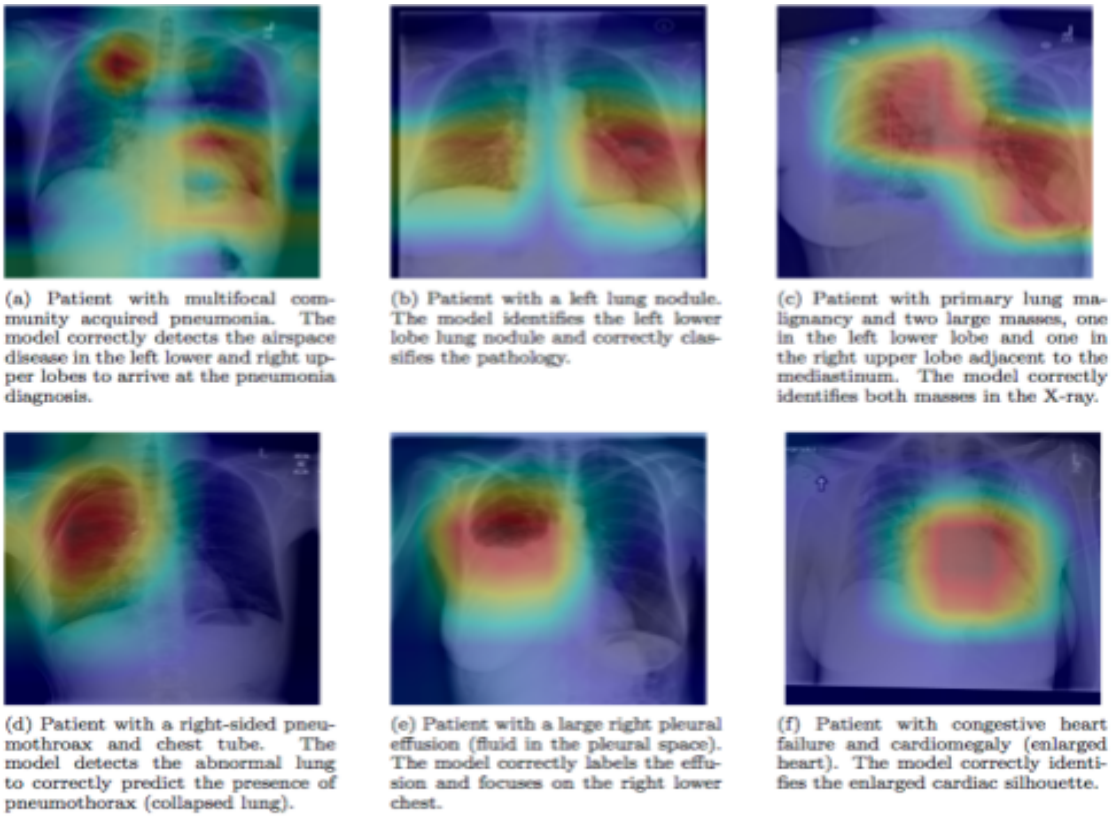
Figure 1.4 Gartner Hype Cycle,Source:
https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/

One of the first concerns mentioned, regarding the use of AI in the health care sector is that AI will replace current professionals, decreasing the overall number of jobs. (Obermeyer & Emanuel, 2016) Should we note that the ability to transform data into knowledge will affect at least three areas of medicine: (i) improving prognosis (ii) displacing the work of radiologists and anatomical pathologists and (iii) improving diagnostic accuracy.

Even though some welcome this technology ,they also mention that the complexity of medicine exceeds the current capacity of the human mind. They argue that medical decisions have become irritably complex and that doctors are not able to reliably evaluate the patients as much as they would hope, leading to dissatisfaction and burnout among doctors (Obermeyer & Lee, 2017).

Even though the replacement of medical doctors with AI has yet to be an immediate and direct issue, it may be beneficial for these specialties most at risk of automation to begin strategically planning for the future in which AI is a part of the healthcare workforce. A role of an “information specialist” is suggested (Jha, n.d.) in which their responsibility would be less focused on the extraction of information from images and histology, but rather shifted to focus on management and interpretation of the information extracted by AI. Regardless of displacement, AI in medicine should be considered an asset, as part of the collective team.

Current prognostic models are restricted only to a few variables, but AI will be able to use far more variables. In addition to handling far larger and more complex data than humans, AI offers several advantages for radiology and other medical image domains: (i) rules are created by the algorithm, which may pick up on a connection, a physician may not have been trained to notice (ii) AI combines predictors in interactive and nonlinear ways (iii) AI is not affected by human emotions, fatigue, or distractions and (iv) AI will increase dramatically over time with the inclusion of larger datasets, more experience, and greater computing power (Chockley & Emanuel, 2016). These systems, providing they are used responsibly (in addition to medical professionals), have the potential to greatly improve diagnosis rates.



(a) Patient with multifocal community acquired pneumonia. The model correctly detects the airspace disease in the left lower and right upper lobes to arrive at the pneumonia diagnosis.

(b) Patient with a left lung nodule. The model identifies the left lower lobe lung nodule and correctly classifies the pathology.

(c) Patient with primary lung malignancy and two large masses, one in the left lower lobe and one in the right upper lobe adjacent to the mediastinum. The model correctly identifies both masses in the X-ray.

(d) Patient with a right-sided pneumothroax and chest tube. The model detects the abnormal lung to correctly predict the presence of pneumothorax (collapsed lung).

(e) Patient with a large right pleural effusion (fluid in the pleural space). The model correctly labels the effusion and focuses on the right lower chest.

(f) Patient with congestive heart failure and cardiomegaly (enlarged heart). The model correctly identifies the enlarged cardiac silhouette.

Figure 1.5 heatmap-like figure of which pixels are considered most important for pneumonia detection, Source: (Rajpurkar et al., 2017)

1.4 Medical Image Analysis

In this work, when discussing medical image analysis, we are not concerned with the physics of imaging, the image instrumentation, or the image acquisition and reconstruction process. Rather, we are concerned with the analysis, modeling, and application of the image in addition to the knowledge base and physiology of the domain. Though advances in medical image analysis have led to many algorithms in commercial settings, that solve medical image analysis tasks with sufficient performance (accuracy, reliability, speed), there still exist challenges that motivate the ongoing development of more accurate, reliable, and faster algorithms that are dedicated to specific applications (Weese & Lorenz, 2016).

1.5 Overview Of Thesis

This thesis is structured as follows: Chapter 2 presents state of the art related work in the areas of dermatology and skin lesion classification, with hand crafted methods. Chapter 3 presents state of the art work in machine learning and deep learning fields respectively, and makes further explanation of CNN, neural networks principles and introduces concepts such as fine tuning, and validation measures.

Chapter 4 makes an introduction to the section of transfer learning, presented with a brief explanation, and the pre-trained networks that we used for our thesis problem.

Chapter 5 presents our experimental framework and results with further explanation and discussion over the results, and also presents the datasets that we used for our problem along with related datasets and challenges on the skin lesion classification section. In the end of the chapter we make an overall comparison of our methods.

Chapter 6 recommends an alternative method of approach, with skin lesion segmentation and the second round of experiments, along with state of the art works and a final discussion and comparison over the results as a conclusion.

Chapter 7 contains concluding remarks and suggestions for future work.

Chapter 2
State Of The Art Techniques And Problems

2.1 Dermoscopy

Despite there being no definitive non-invasive method of diagnosing melanoma, capturing and analyzing an image of a patient's skin is a common non-invasive method of attempting to diagnose a suspicious skin lesion (Goodson & Grossman, 2009). These images can either be macroscopic or dermoscopic. Dermoscopic images are produced by using specialty equipment such as cross-polarizing light filters (non-contact dermoscopy) or an oil/gel interface (immersion contact dermoscopy) (Goodson & Grossman, 2009). Macroscopic images are considered images captured by more conventional systems, such as a standard camera. Regarding a comparison of clinical photography and dermoscopy, please see Figure 2.4. The difference when comparing dermoscopic images vs macroscopic images are quality and cost. Dermoscopic images provide additional color and pattern properties (Menzies & Zalaudek, 2006) but may require more effort and cost to acquire, whereas macroscopic images typically require a less complex and cheaper alternative at the loss of image details. Dermoscopy has been shown to increase diagnosis accuracy [(Argenziano et al., 2007), (Bafounta et al., 2001)] compared to naked-eye examination. Though the techniques discussed can be applied to either image type (dermoscopic or macroscopic), due to dataset availability (see Section 2.3.4) and adaption of dermoscopy (Murzaku et al., 2014), we will be referring to dermoscopic images in our work.

2.2 Conventional Methods

Unaided Methods Dermatologists have developed many methods relying on dermoscopic criteria to aid in reducing the subjectivity of determining whether a skin lesion is malignant. Some focus only on recognizing melanoma related criteria (such as the Menzies method (Menzies et al., 2003) and the 7-point checklist (Argenziano et al., 1998). Others, like pattern analysis (Pehamberger et al., 1987), focus on the identification and their density inside the skin lesions. Yet another group combines

the dermoscopic criteria and a high-level analysis of the lesions (such as analyzing the overall shape, size, and border of the lesion as well).

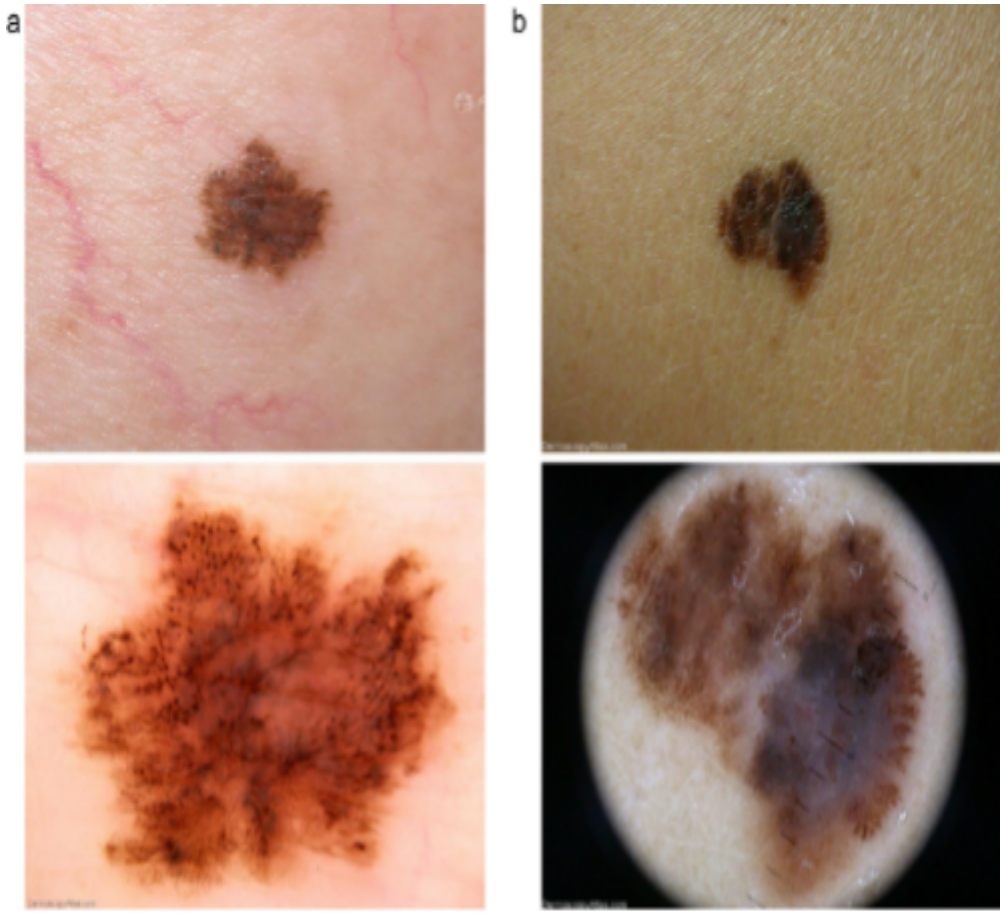


Figure 2.1:Images of pigmented skin lesions by clinical photography (top) and dermoscopy (bottom): (a) In situ melanoma and (b) invasive melanoma. Used with permission. , submitted by Dr. Alan Cameron (a) and Dr. Jean-Yves Gourhant (b). Source:https://www.researchgate.net/figure/Images-of-pigmented-skin-lesions-by-clinical-photography-top-and-dermoscopy-bottom_fig2_232249318

Two of the most well known methods are the ABCD (Stolz et al., 1991) (later ABCDE (Rigel et al., 2005)) and CASH (Henning et al., 2007). ABCD rule (Asymmetry, Border, Color, and Differential Structure) and CASH algorithm (Color, Architecture, Symmetry, Homogeneity).

A skin lesion examined using the ABCD method may be evaluated as follows:

• Asymmetry – a lesion is first assigned a major and minor axes and the opposite halves are compared in terms of shape, color, and pattern, where a lesion may receive: i) 0 points (fully symmetric), ii) 1 point (asymmetric on one axis), or ii) 2 points (asymmetric on both axes).

•Border – a lesion is first divided into eight imaginary slices by a dermatologist. For each slice, a score is assigned depending on how abrupt the transition from lesion to surrounding skin is. An abrupt transition receives a score of 1, otherwise a score of 0 is assigned.

• Color Features – a lesion points corresponding to the number of
colors measured are assigned (white, red, light brown, dark brown, blue-gray, and black). A minimum of 1 is possible, while the highest possible is 6.

• Differential structures – a lesion is assigned points for every differential structure observed (homogeneous areas, dots, globules, and streaks).

To calculate a final score, each individual score is multiplied by a weight factor and totaled to give a final value. The value is then compared to a threshold value and a classification is assigned. For more information on these weight factors and threshold, please see Nashbar et al. (Nachbar et al., 1994).

Examining the differences between skin lesions in photographs is non-trivial, even to a trained medical doctor. Due to the subjectivity and less than ideal performance attainable by professionals, a strong motivation exists to develop and investigate the use of computer aided diagnosis (CADx) systems, which might improve the accuracy and sensitivity of melanoma detection methods.

As we can see in the image below, the differences between the lesions are difficult to detect without a dermatoscope and the doctor's knowledge.

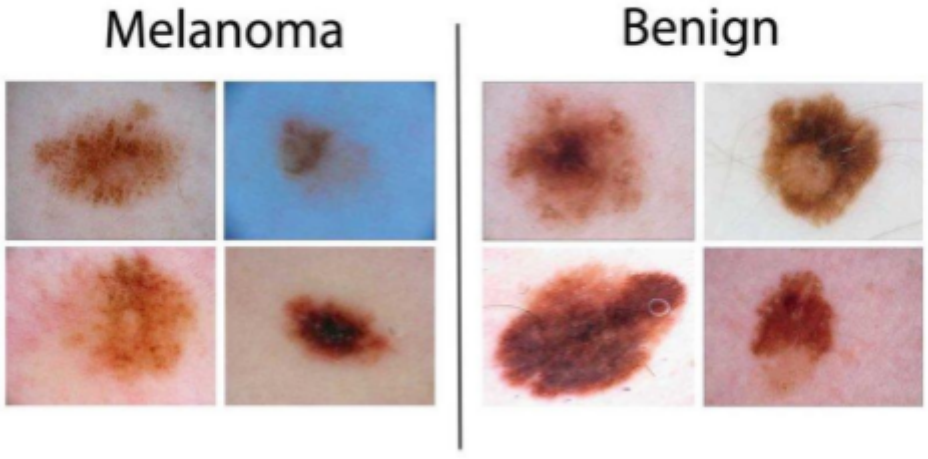


Figure 2.2:Sample images created from the ISIC Archive dataset
Source:https://www.researchgate.net/figure/Sample-images-created-from-the-ISIC-Archive-dataset-5_fig1_314203318

2.2.1 Feature Extraction combined with lesion classifiers

In this section, the feature extraction and classification components from the classical machine learning pipeline (Figure 2.3) are discussed. Many classical malignant classification methods consist of 3-4 consecutive steps, usually inspired by the ABCD rule, and often rely on hand-crafted features, such as: lesion type (primary morphology), lesion configuration (secondary morphology), color, distribution, shape, texture, and border irregularity measures (Page, n.d.).

The feature extraction phase is often performed in an attempt to imitate performance of dermatologists by extracting dermoscopic structures like: irregular streaks and regression structures (Fabbrocini et al., 2010), pigment network [(Barata et al., 2011),(Caputo et al., 2002)(Di Leo et al., 2008), (Fleming et al., 1998), (Sadeghi et al., 2011)], granulari- ties (Stoecker et al., 2011), blue-white veil (Celebi et al., 2008), dots (Yoshino et al., 2004), globules (Fleming et al., 1998), blotches [(Pellacani et al., 2004),(Stoecker et al., 2005)]. Additional works attempt to isolate and evaluate features such as pigment distribution [(Seidenari et al., 2005)], relative color histograms [(Faziloglu et al., 2003), (Stanley et al., 2007)], and texture descriptors [(Sheha et al., 2012), (Yuan et al., 2006)]. Systems have even been developed to mimic unaided methods such as the seven-point checklist [(Di Leo et al., 2010)].

Though these works achieve promising results, the difficulty of these methods is that they require hand-coded, low-level features, that may even be based on subjective thresholds/values. Additionally, these methods typically require the lesion to be isolated (segmented) from artifacts. This creates a scaling issue. Creating these features is time consuming, requires expertise, and may not generalize to larger, more diverse, datasets.

2.3 Image Processing and Feature Extraction

At first, we tried to solve our problem by isolating the skin and nevus area in our images from background and hair.

2.3.1 Skin Detection

Initially we tried to make a reliable skin detection algorithm, by applying thresholds on the images after we converted them first in the CIELab color space. Afterwards,

we tried to find the best spaces using trackbars by opencv, having as base the prices from the paper Machine Learning and Knowledge Extraction (Holzinger et al., 2019,) and applying the skin masks at the images.

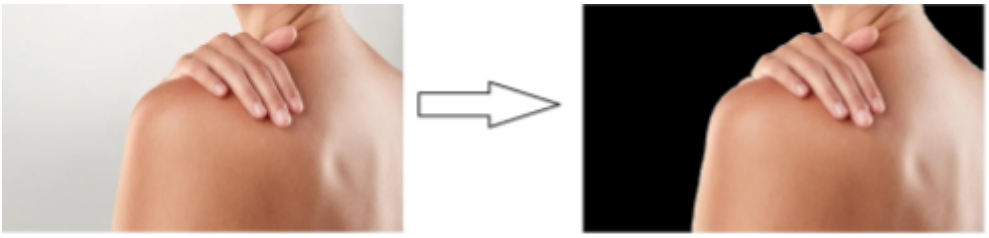


Figure 2.3 Skin Detection Using CIELab Color space

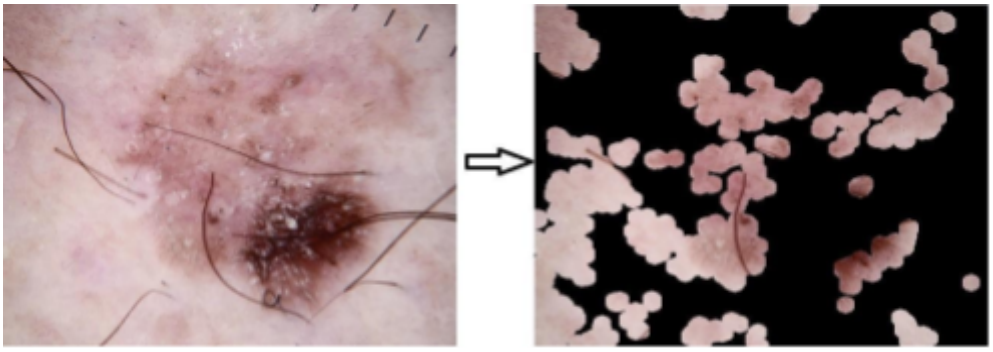


Figure 2.4 Skin Detection Using CIELab on ISIC Image

2.3.2 Hair Removal

In the hair removal problem, we tried to find a solution by following some specific steps:

1.Convert the color image to a grayscale version.

2.Applying Morphological Black-Hat transformation on the grayscale image.

3.Creating the mask for the InPainting task.

4.Applying inpainting algorithm on the original image using the mask prepared from the grayscale image in step 3.

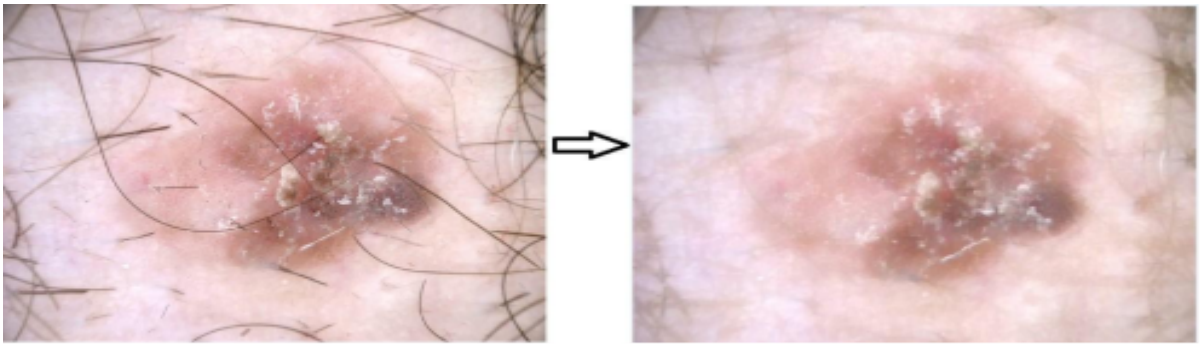


Figure 2.5 Hair Removal on ISIC Image

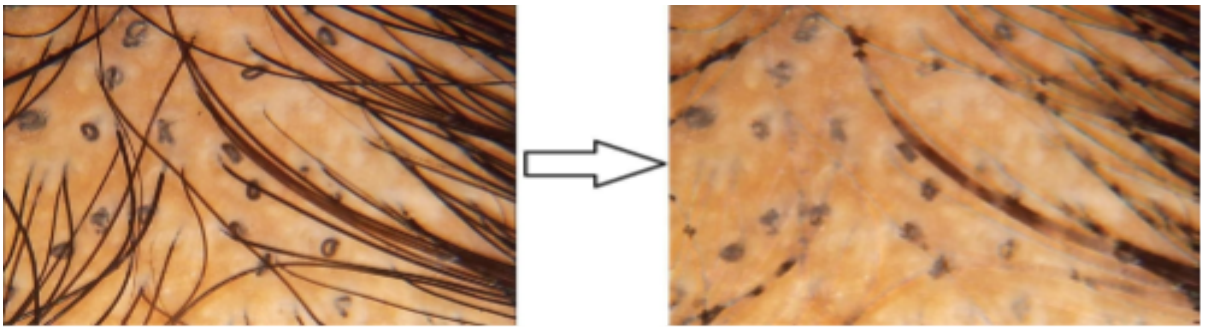


Figure 2.6 Hair Removal on Internet Image

2.3.4 Disadvantages of hair removal and skin detection

The problem, regarding these approaches, is that in the skin detection algorithm we have different estimations that depend on the image morphology and factors such as contrast, illumination etc. In figure 2.3 we see a good skin detection while we want to keep the skin area and remove the background , although in figure 2.4 we see that the algorithm blackens the nevus area which is not acceptable as a reliable result. Also in the hair removal problem we see that the algorithm in figure 2.5, apparently removes the hair but when we zoom the image we notice that the image is blurred, while in figure 2.6, we also do not have a good result as many of the hair are still intact. So in order to have more efficient algorithms and better results, (of course because the company's program works specifically with dermoscopic images), we work with corresponding datasets and machine and deep learning methods.

Chapter 3
An Introduction To Machine And Deep Learning Methods

As we proceeded through experiments for this thesis, we realised that hand crafted methods come along with limitations about the solutions for our problem. Limitations about the skin tone and different conditions of lumination, make the reliability of hand crafted methods low enough to search for other methods.
Therefore we moved towards the artificial intelligence section and especially in its subsets, machine learning and deep learning.

3.1 Machine Learning

The scientific field of machine learning (ML) is a branch of artificial intelligence, as defined by Computer Scientist and machine learning pioneer Tom M. Mitchell: "Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience."
An algorithm can be thought of as a set of rules/instructions that a computer programmer specifies, which a computer can process. Simply put, machine learning algorithms learn by experience, similar to how humans do. Machine Learning is used anywhere from automating mundane tasks to offering intelligent insights, industries in every sector are trying to benefit from it. For example, image and speech recognition, financial industry and trading and most important in medical diagnoses.

3.1.1 Introduction in Machine Learning

Process:

1. Data Collection: Collect the data that the algorithm will learn from.

2. Data Preparation: Format and engineer the data into the optimal format, extracting important features (feature extraction) and performing

dimensionality resizing.

3. Training: This is where the machine learning algorithm actually learns by showing it the data that has been collected and prepared.

4. Evaluation: Test the model to see how well it performs.

5. Tuning: Fine tune the model to maximise its performance.

Machine learning algorithms can be classified into three key categories based on the different types of learning problems addressed. A list of these categories is:

- Supervised Learning: In supervised learning, the training dataset needs to be in a specific format. Each instance (data point) has an assigned label. Datasets are labeled as $(x, y) \in X \times Y$, where x and y denote a data point and the corresponding true prediction for x. If the output y is part of a discrete domain, the problem is a classification task. If the output belongs to a continuous domain, then it is a regression task.

- Unsupervised Learning: Unlike supervised learning, the datasets are not labeled in unsupervised learning. In order to develop a structure from unlabeled data, the ML algorithm should examine the similarities between object pairs.

- Semi-supervised Learning: This learning task is a class of supervised learning and uses a large amount of unlabeled data for training along with the small amount of labeled data.

As of machine learning approaches in the field of dermoscopy, (Ali et al., 2020) reports state-of-the-art performance by proposing an automated approach for skin lesion border irregularity detection.
The approach involves extracting the skin lesion from the image, detecting the skin lesion border, measuring the border irregularity, training a Convolutional Neural Network and Gaussian naive Bayes ensemble, to the automatic detection of border irregularity, which results in an objective decision on whether the skin lesion border is considered regular or irregular. The approach achieves outstanding results,

obtaining an accuracy, sensitivity, specificity, and F-score of 93.6%, 100%, 92.5% and 96.1%, respectively.

(Ozkan & Koklu, 2017) proposed a method in order to pre-classify the skin lesions in three groups as normal, abnormal and melanoma by machine learning methods and to develop a decision support system that should make the decision easier for a doctor. The objective of this study is skin lesions based on dermoscopic images PH2 datasets using 4 different machine learning methods namely; ANN, SVM, KNN and Decision Tree. Correctly classified instances were found as 92.50%, 89.50%, 82.00% and 90.00% for ANN, SVM, KNN and DT respectively. The findings show that the system developed in this study has the feature of a medical decision support system which can help dermatologists in diagnosing skin lesions.

After feature extraction, machine learning methods, such as k-nearest neighbors (kNN), Artificial Neural Networks (ANNs), the multilayer feedforward network [(Rumelhart et al., 1987)], logistic regression [(Menard, 2018)], decision trees [(Safavian & Landgrebe, 1991)] and support vector machines (SVMs) [(Burges, 1998 (Safavian & Landgrebe, 1991))], are employed to perform classification. Moderate success has been achieved with these methods [(Dreiseitl et al., 2001)]. Some examples of related work using this classification pipeline (hand-crafted features and popular classifiers) are mentioned below. Barata et al. [(Barata et al., 2014)], notes these above mentioned implementations and compares global and local features against three classifiers AdaBoost [(Freund & Schapire, 1997)], SVM, and kNN.

They report that color features perform much better than texture features alone and that both global and local features achieve promising results (slight advantage to local). Other works implement methodologies combining feature extraction methods and using classifiers such as logistic regression [(Blum et al., 2004)] and SVMs [(Yuan et al., 2006)]. Recently, ensemble methods have been proposed which further examine the combination of feature extractors and classifiers [(Rastgoo et al., 2015), (Schaefer et al., 2014)]. There has also been work as early as 1994 which proposed using neural networks to classify skin lesions [(Binder et al., 1994), (Ercal et al., 1994), (Piccolo et al., 2002)].

3.2 Deep Learning

3.2.1 Introduction in DL

Deep learning, as the name suggests, is a subset of machine learning. Deep Learning mostly involves using deep artificial neural networks (algorithms/computational models loosely inspired by the human brain) to tackle machine learning problems.

3.2.2 State of the art (DL)

Recently, the emergence of deep learning has led to the development of promising classification methodologies. More specifically, convolutional neural networks (CNNs) (LeCun et al., 1998) have achieved promising results classifying skin lesions, capable of potentially outperforming medical professionals working on the same task [(Codella et al., 2015),(Esteva et al., 2017)]. In a conventional setting, feature extraction (often with handcrafted features) is performed before being classified by a classifier (such as an MLP or SVM).

When performing the same classification with deep learning, the feature extraction and classification are both learned and performed as a single unit (See Figure 2.6). Recent Implementations Codella et al. explore using an SVM trained on features extracted from CNN (Jia et al., 2014) model pretrained on the ILSVRC 2012 dataset (Krizhevsky et al., 2012)) using transfer learning and report performance on par with ensembles of low-level features (Codella et al., 2015). (Kawahara et al., 2016) explores the possibility of using a pretrained CNN as a feature extractor, rather than training the architecture from scratch. Furthermore, their work demonstrates that using the learned filters from a CNN pretrained on natural images generalize to classifying 10 classes of non-dermoscopic skin images.

Fine tuning ImageNet (Deng et al., 2009) pretrained models were performed by Laio (Liao, 2016) in an attempt to create a universal skin disease classification system. (Codella et al., 2016) reports state-of-the-art performance using an ensemble of methods including low level features like color, edge and multiscale local binary patterns, with sparse coding (gray and RGB), pretrained-pre-trained model from the Image Large Scale Visual Recognition Challenge (ILSVRC) 2012 dataset (Russakovsky et al., 2015) and a Deep Residual Network (DRN) (He et al., 2015). (Kawahara & Hamarneh, 2016), proposed a multi-resolution-tract CNN.

Though they are met with hardware constraints, the work proposes using multiple resolution inputs to classify skin lesions where lower resolutions may provide high level features like overall shape and high resolution inputs may provide additional low level features like patterns within the lesion. (Esteva et al., 2017) fine tune Inceptionv3 and benchmark their results on a large, proprietary dataset, against 21 board-certified dermatologists.
Their findings indicate that their solution is capable of classifying skin lesions to a level comparable to trained dermatologists. Result summaries from the ISBI challenges (introduced in Section 2.3.4) have also been made public. The results and entries of the 2016 Challenge (Gutman et al., 2016) are discussed in (Marchetti et al., 2018) and the summary of the 2017 Challenge Summary is outlined in (Codella et al., 2017). A review of current state-of-the-art methodologies is presented in (Pathan et al., 2018).

3.3 Differences between Machine and Deep Learning

In practical terms, deep learning is just a subset of machine learning. In fact, deep learning technically is machine learning and functions in a similar way (hence why the terms are sometimes loosely interchanged). However, its capabilities are different.
While basic machine learning models do become progressively better at whatever their function is, they still need some guidance. If an AI algorithm returns an inaccurate prediction, then an engineer has to step in and make adjustments. With a deep learning model, an algorithm can determine on its own if a prediction is accurate or not through its own neural network.

To recap the differences between the two:

- Machine learning uses algorithms to parse data, learn from that data and is based on feature extraction and clustering/classification (decision making). Also machine learning algorithms make informed decisions based on what they have learned and work with regression algorithms or decision trees.

- Deep learning structures algorithms in layers to create an "artificial neural network" that can learn and make intelligent decisions on its own and uses

neural networks that function very similarly to the biological neural connections of our brain.

- Deep learning is a subfield of machine learning. While both fall under the broad category of artificial intelligence, deep learning is what powers the most human-like artificial intelligence.

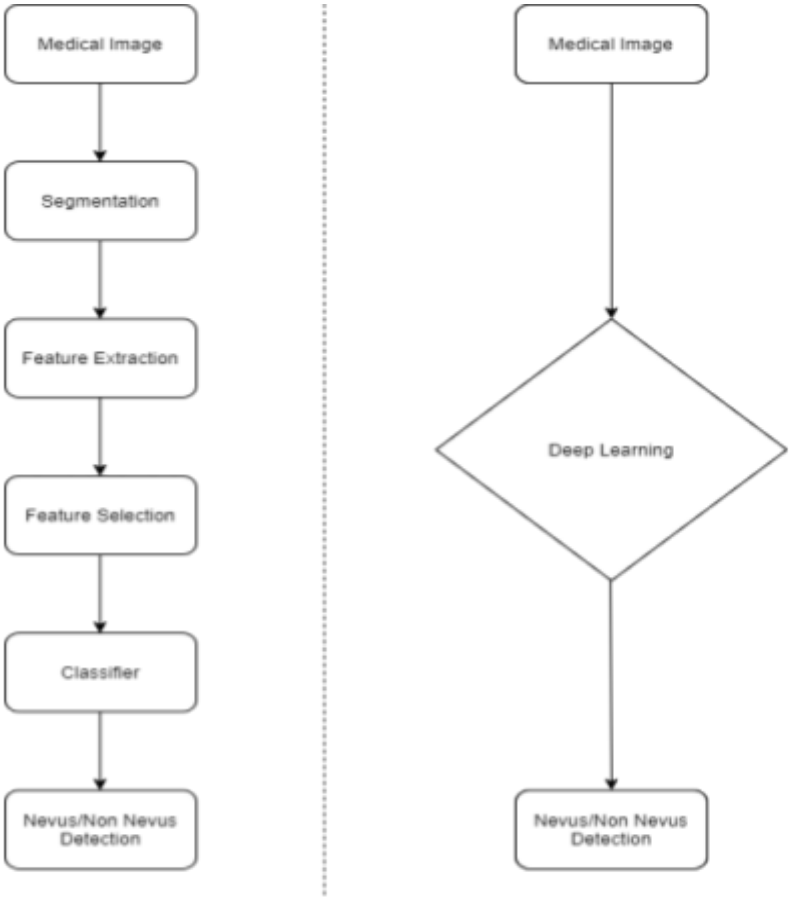


Figure 3.1 :Distinction between an advanced classical machine learning pipeline and
a deep learning pipeline to perform classification of skin lesion images

3.4 Medical Image Analysis using Machine and Deep Learning

Deep Learning, a subset of machine learning which is a subset of artificial intelligence, encompasses many different principles and domains. Recently, the performance and adaptation of deep learning based approaches to medical image analysis has increased performance dramatically, to the point of raising concerns around the future of current professions; such as the human radiologist [(Jha, 2016), (Walter, 2016)]. Figure 2.3 demonstrates how deep learning has effectively merged and replaced entire components of a classical machine learning pipeline for image analysis in a medical setting. In a classical setting, images typically undergo:

• Segmentation – Where unnecessary information is removed from the image
• Feature Extraction – Where hand-crafted features, dictionary-based features, and or clinically inspired features are extracted from the image
• Feature Selection – Where the most relevant features are selected
• Classification – Where predicted label(s) are generated for the image

Many of the above listed components need to be performed as separate ML tasks (segmentation) or even by hand (selecting which features to extract). In a deep learning pipeline, it is not uncommon to have the deep learning architecture completely replaced all these components with one deep learning architecture that performs all stages in a supervised setting. Litjens et al. provide a recent survey on the use of deep learning in medical image analysis (Litjens et al., 2017)

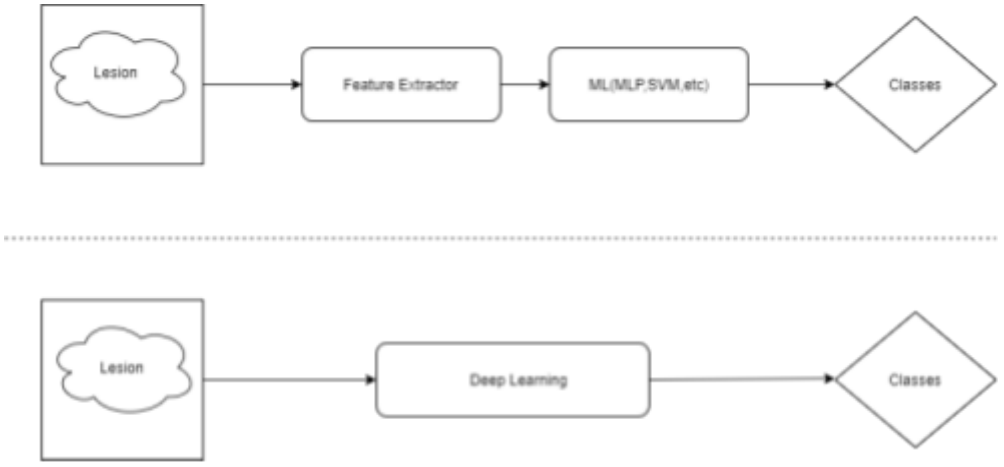


Figure 3.2:Distinction between a classical skin lesion classification pipeline using pre-segmented dermoscopic images in a conventional ML pipeline vs a DL pipeline.

3.5 Convolutional Neural Network

The basic unit of computation in a neural network is the neuron which is often called a node or unit. It receives input from some other nodes, or from an external source and computes an output. Each input has an associated weight (w), which is assigned on the basis of its relative importance to other inputs. The node applies a function f (defined below) to the weighted sum of its inputs as in figure below.

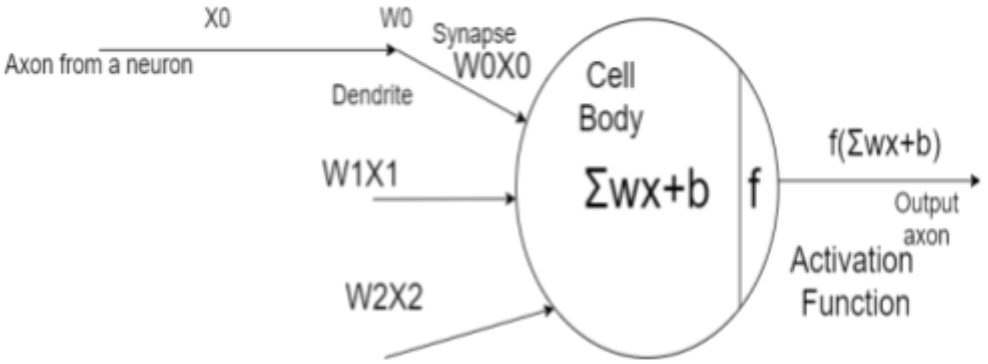


Figure 3.3 Mathematical Model of a single neuron in a neural network

The above network takes numerical inputs X1 and X2 and has weights w1 and w2 associated with those inputs. Additionally, there is another input 1 with weight b (called the Bias) associated with it.A Convolutional Neural Network (CNN) is a deep learning algorithm that can recognize and classify features in images for computer vision. It is a multi-layer neural network designed to analyze visual inputs and perform tasks such as image classification, segmentation and object detection, which can be useful for autonomous vehicles. CNNs can also be used for deep learning applications in healthcare, such as medical imaging. There are two main parts to a CNN:

- A convolution tool that splits the various features of the image for analysis
- A fully connected layer that uses the output of the convolution layer to predict the best description for the image.

CNN architecture is inspired by the organization and functionality of the visual cortex and designed to mimic the connectivity pattern of neurons within the human brain. The neurons within a CNN are split into a three-dimensional structure, with each set of neurons analyzing a small region or feature of the image. In other words, each group of neurons specializes in identifying one part of the image. CNNs use the

predictions from the layers to produce a final output that presents a vector of probability scores to represent the likelihood that a specific feature belongs to a certain class.

3.5.1 The CNN layers

A CNN is composed of several kinds of layers:

- Convolutional layer—creates a feature map to predict the class probabilities for each feature by applying a filter that scans the whole image, few pixels at a time.
- Pooling layer (downsampling)—scales down the amount of information the convolutional layer generated for each feature and maintains the most essential information (the process of the convolutional and pooling layers usually repeats several times).
- Fully connected input layer—"flattens" the outputs generated by previous layers to turn them into a single vector that can be used as an input for the next layer.
- Fully connected layer—applies weights over the input generated by the feature analysis to predict an accurate label.
- Fully connected output layer— generates the final probabilities to determine a class for the image.

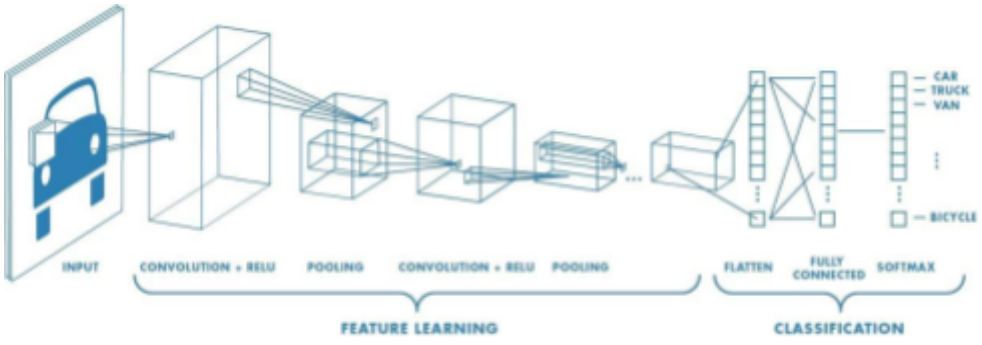


Figure 3.4 CNN Example
Source:https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

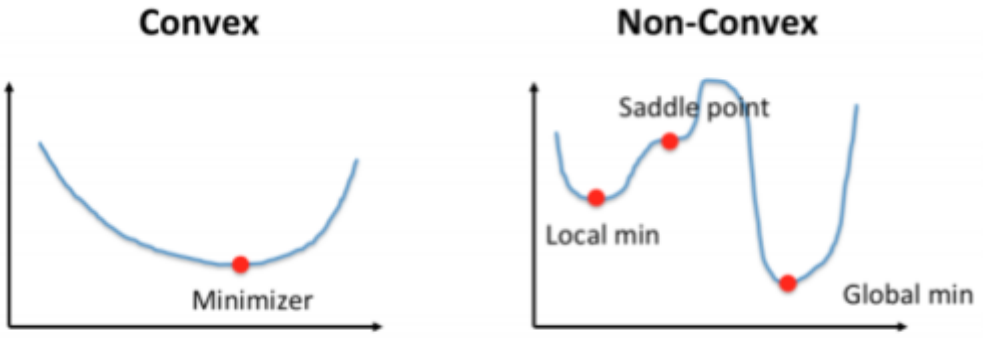
Simply put, an activation function is a function that is added into an artificial neural network in order to help the network learn complex patterns in the data. When comparing with a neuron-based model that is in our brains, the activation function is at the end deciding what is to be fired to the next neuron. That is exactly what an activation function does in an ANN as well. It takes in the output signal from the previous cell and converts it into some form that can be taken as input to the next cell. Various nonlinear activation functions have been used, such as the sigmoid, Softmax and ReLU (Rectified Linear Unit).
Forward propagation is how neural networks make predictions. Input data is “forward propagated” through the network layer by layer to the final layer which outputs a prediction.
Backpropagation, short for "backward propagation of errors", is an algorithm for supervised learning of artificial neural networks using gradient descent. Given an artificial neural network and an error function, the method calculates the gradient of the error function with respect to the neural network's weights.

3.5.2 Optimization

Optimization refers to the task of minimizing/maximizing an objective function f(x) parameterized by x. In machine/deep learning terminology, it’s the task of minimizing the cost/loss function J(w) parameterized by the model’s parameters w ∈ R^d.
Optimization algorithms (in the case of minimization) have one of the following goals:

1.Find the global minimum of the objective function. This is feasible if the objective function is convex, i.e. any local minimum is a global minimum.

2.Find the lowest possible value of the objective function within its neighborhood. That’s usually the case if the objective function is not convex as the case in most deep learning problems.
Gradient Descent is an optimizing algorithm used in Machine/ Deep Learning algorithms. The goal of Gradient Descent is to minimize the objective convex function f(x) using iteration.



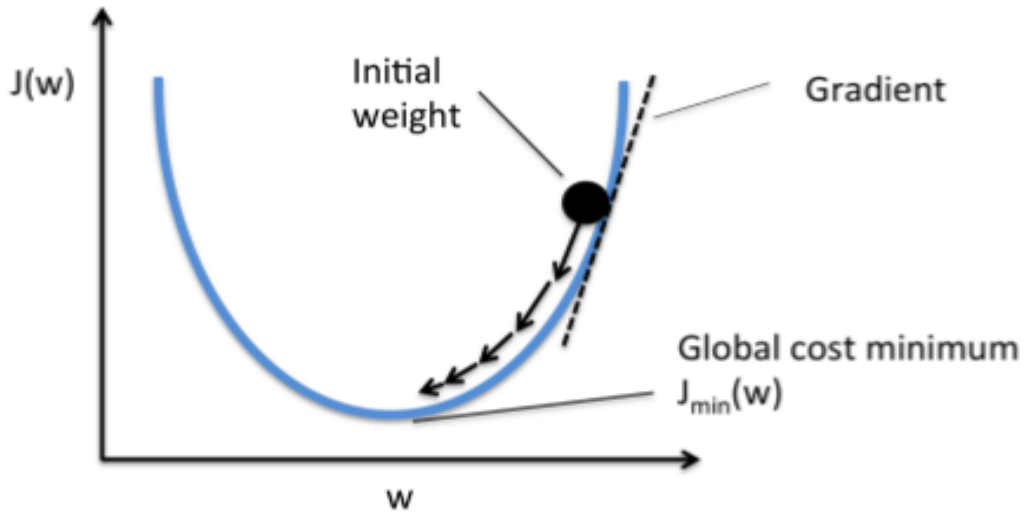


Figure 3.6 Intuition behind Gradient Descent,
Source:https://hackernoon.com/gradient-descent-aynk-7cbe95a778da

A machine learning model always wants low error with maximum accuracy, in order to decrease error we will intuit our algorithm that you're doing something wrong that needs to be rectified, that would be done through Gradient Descent.

3.5.3 Model Training, Fine-tuning, and Validation

Model training is performed on the training set and validated on a separate validation set, which is not used for training purposes (i.e. k-fold cross validation (Kohavi, 1995). This happens because the kaggle dataset that we used in our case, has already separated sets as training, test and validation.
Afterward for reasons of completeness we perform a 10-fold cross validation.
The output of a specified loss function, which represents the difference between the ground truth and the prediction, is optimized during training to be the lowest value (smallest difference between the ground truth and predicted value). The term cost is frequently used interchangeably with loss. Technically, the loss refers to the error on a single example, whereas the cost is the average of the loss across the entire training set. In this work, the term loss is used exclusively and refers to the average loss value over a specified batch or epoch. As shown in Figure 3.7, training is considered finished when the validation loss no longer improves. The network parameters

(weights and biases) are then saved when the validation loss is recorded at its lowest.

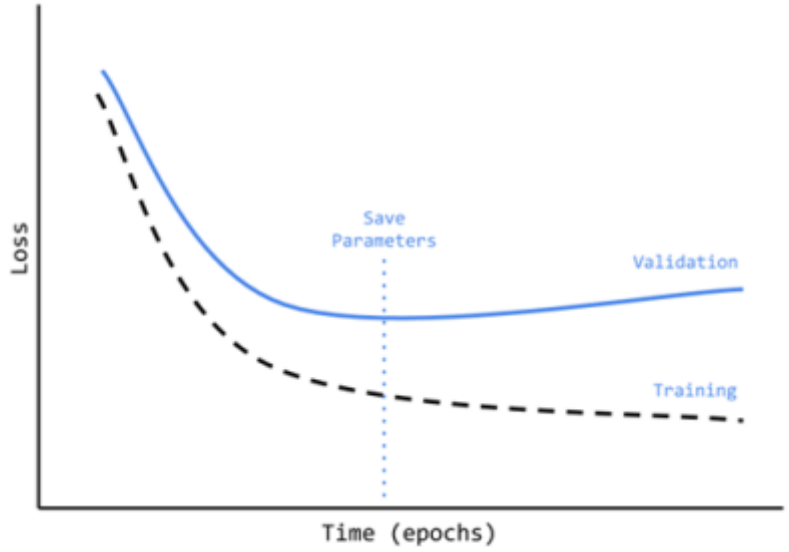


Figure 3.7 Example loss plot showing possible training and validation loss

When a metric does not improve further, we decrease learning rate in order to avoid overfitting.

Early Stopping demands from a validation dataset to be evaluated during training, and it stops the training process after a certain number of epochs insomuch the metric of our interest does not improve.

3.5.4 The Vanishing/Exploding Gradient Problem in DL

When training a deep neural network with gradient based learning and backpropagation, we find the partial derivatives by traversing the network from the final layer (y_hat) to the initial layer. Using the chain rule, layers that are deeper into the network go through continuous matrix multiplications in order to compute their derivatives.

In a network of n hidden layers, n derivatives will be multiplied together. If the derivatives are large then the gradient will increase exponentially as we propagate down the model until they eventually explode, and this is what we call the exploding gradient problem. Alternatively, if the derivatives are small then the gradient will decrease exponentially as we propagate through the model until it eventually vanishes, and this is the vanishing gradient problem.

In the case of exploding gradients, the accumulation of large derivatives results in the

model being very unstable and incapable of effective learning. The large changes in the models weights create a very unstable network, which at extreme values the weights become so large that it causes overflow resulting in NaN weight values of which can no longer be updated. On the other hand, the accumulation of small gradients results in a model that is incapable of learning meaningful insights since the weights and biases of the initial layers, which tends to learn the core features from the input data (X), will not be updated effectively. In the worst case scenario the gradient will be 0 which in turn will stop the network will stop further training.

3.5.5 Batch Normalization in Neural Networks

We normalize the input layer by adjusting and scaling the activations. For example, when we have features from 0 to 1 and some from 1 to 1000, we should normalize them to speed up learning. If the input layer is benefiting from it, why not do the same thing also for the values in the hidden layers that are changing all the time and get 10 times or more improvement in the training speed.
Batch normalization reduces the amount by what the hidden unit values shift around (covariance shift). To explain covariance shift, let's have a deep network on cat detection. We train our data on only black cats' images. So, if we now try to apply this network to data with colored cats, it is obvious; we're not going to do well. The training set and the prediction set are both cats' images but they differ a little bit. In other words, if an algorithm learned some X to Y mapping, and if the distribution of X changes, then we might need to retrain the learning algorithm by trying to align the distribution of X with the distribution of Y. Also, batch normalization allows each layer of a network to learn by itself a little bit more independently of other layers.
To increase the stability of a neural network, batch normalization normalizes the output of a previous activation layer by subtracting the batch mean and dividing by the batch standard deviation. However, after this shift/scale of activation outputs by some randomly initialized parameters, the weights in the next layer are no longer optimal. SGD (Stochastic gradient descent) undoes this normalization if it's a way for it to minimize the loss function. (D, 2020)

3.5.6 Dropout

The term "dropout" refers to dropping out units (both hidden and visible) in a neural network.Simply put, dropout refers to ignoring units (i.e. neurons) during the training

phase of a certain set of neurons which is chosen at random. By “ignoring”, I mean these units are not considered during a particular forward or backward pass.More technically, at each training stage, individual nodes are either dropped out of the net with probability 1-p or kept with probability p, so that a reduced network is left; incoming and outgoing edges to a dropped-out node are also removed.

Why do we need Dropout?

Given that we know a bit about dropout, a question arises — why do we need dropout at all? Why do we need to literally shut-down parts of neural networks?The answer to these questions is “to prevent over-fitting”.A fully connected layer occupies most of the parameters, and hence, neurons develop co-dependency amongst each other during training which curbs the individual power of each neuron leading to over-fitting of training data.Now that we know a little bit about dropout and the motivation, let’s go into some more details. If you just wanted an overview of dropout in neural networks, the above two sections would be sufficient. In this section, I will touch upon some more technicality.In machine learning, regularization is a way to prevent over-fitting. Regularization reduces overfitting by adding a penalty to the loss function. By adding this penalty, the model is trained such that it does not learn an interdependent set of features weights. Those of you who know Logistic Regression might be familiar with L1 (Laplacian) and L2 (Gaussian) penalties.Dropout is an approach to regularization in neural networks which helps reduce interdependent learning amongst the neurons.

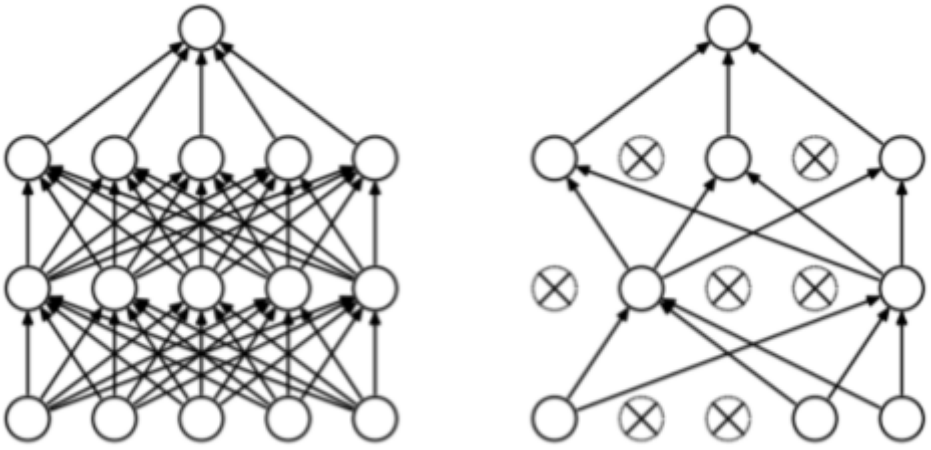


Figure 3.8 Dropout, Source :https://medium.com/@amarbudhiraja/https-medium-com-amarbudhiraja-learning-less-to-learn-better-dropout-in-deep-machine-learning-74334da4bfc5

Training Phase:
For each hidden layer, for each training sample, for each iteration, ignore (zero out) a random fraction, p, of nodes (and corresponding activations).

Testing Phase:
Use all activations, but reduce them by a factor p (to account for the missing activations during training).
In the following tasks, we have used the dropout method with good performance.

3.6 Validation Measures

There are various metrics which we can use to evaluate the performance of ML algorithms, classification as well as regression algorithms. We must carefully choose the metrics for evaluating ML performance because −

- How the performance of ML algorithms is measured and compared will be dependent entirely on the metric you choose.
- How you weigh the importance of various characteristics in the result will be influenced completely by the metric you choose.

Confusion Matrix

It is the easiest way to measure the performance of a classification problem where the output can be of two or more types of classes. A confusion matrix is nothing but a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, “False Negatives (FN)” as shown below −

		Actual: 1	Actual: 0
Predicted	1	True Positives (TP)	False Positives (FP)
	0	False Negatives(FN)	True Negatives (TN)

Table 3.1 Confusion Matrix

Classification Accuracy

It is the most common performance metric for classification algorithms. It may be defined as the number of correct predictions made as a ratio of all predictions made. We can easily calculate it by confusion matrix with the help of following formula –

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

We can use the accuracy_score function of sklearn.metrics to compute accuracy of our classification model.

Precision

Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula –

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

Recall or Sensitivity

Recall may be defined as the number of positives returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula –

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

Specificity

Specificity, in contrast to recall, may be defined as the number of negatives returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula –

$$Specificity = \frac{TN}{TN + FP}$$

F1 Score

This score will give us the harmonic mean of precision and recall. Mathematically, F1 score is the weighted average of the precision and recall. The best value of F1 would be 1 and worst would be 0. We can calculate F1 score with the help of following formula –

$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

F1 score is having equal relative contribution of precision and recall.

However, in real life problems not all classes have the same importance. It is much more crucial to classify a melanoma as a benign nevi than the opposite.

Chapter 4
Transfer Learning And Deep Learning Structures

4.1 Transfer Learning

As we will see in the Chapter 6, the accuracy of the shallow network is about 83%. However, in biomedicine and dermoscopic applications, this is not a good enough percentage because of the human-level accuracy. If a network's accuracy is near the human level then it is acceptable. In this specific task, the human level accuracy reaches 95%. It is well known that in neural networks, if we want to increase the accuracy we use continuously greater datasets even bigger than 1 million images.
In fact, the datasets that are available for this task are small. In order to solve this problem we use transfer learning, a method widely used in computer vision projects, because it helps us to create high accuracy models in less training time.
First of all, we need to explain what fine tuning is.
We take a model trained over other datasets and use it for our problem. Usually the last layer is a neural network with the same number of units and classes that we want to categorize.
Due to the fact that in our problem we want to categorize 2 classes (nevus or non nevus), we take the network as it is with pretrained weights. Then, we subtract the last layer or the last layers and add a new classifier made by us with 2 units.
Finally, we train our classifiers with our datasets of dermoscopy and this process is called fine tuning.
With the method of transfer learning instead of beginning the training process from the start, we use patterns which have learnt to solve different problems in order to take advantage of the prior knowledge.
As an example, consider the problem of learning how to ride a motorcycle having the knowledge of balance from riding a bicycle.
In that way we do not have to begin from ground zero.
So they use something that they call pre-trained models, i.e. models that are trained in very big datasets for a general purpose (Imagenet is one of them).
They claimed then, that if we take such a model and use the pretrained weights and parameters and train a small amount of layers, and especially the the last ones that

are doing the classification tasks, then we can relate a large amount of information with our datasets which is smaller, and with that method to increase accuracy.
This process is called transfer learning (Pan & Yang, 2010).
Transfer learning is used widely for biomedicine applications and in conjunction with computer vision section, there are remarkable works for melanoma classification (Sagar & J, 2020) or more generally, for skin lesions classification[(Ratul et al., 2019), (Lopez et al., 2017), (Hosny et al., 2019)]. Due to the fact that the total data we worked with is small, which is also true for most of the datasets for similar dermoscopy problems, we use pre trained deep learning structures.

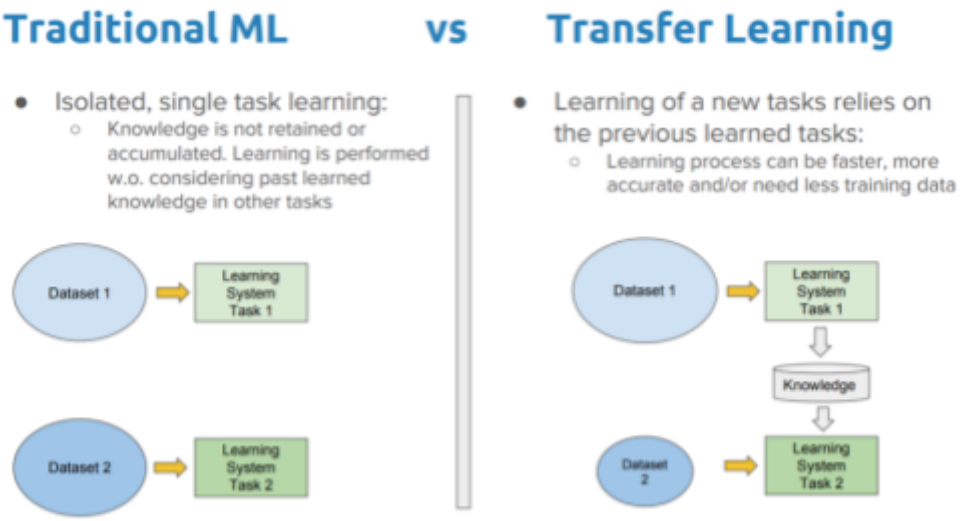


Figure 4.1
Source:https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a

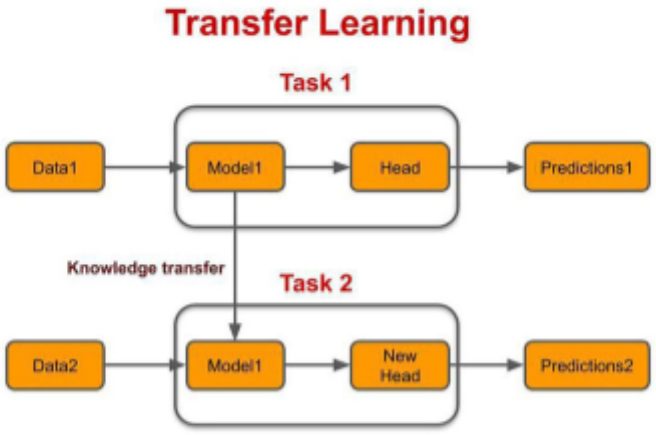


Figure 4.2 Source:https://www.topbots.com/transfer-learning-in-nlp/

4.2 Network Structures

4.2.1 Vgg19

Vgg19 is a simple deep learning structure that takes an RGB image as input and then collects the features hierarchically. It is using only 3×3 convolutional layers stacked on top of each other in increasing depth. Reducing volume size is handled by max pooling. Two fully-connected layers, each with 4,096 nodes are then followed by a softmax classifier (above).

Background

AlexNet came out in 2012 it and improved on the traditional Convolutional neural networks, so we can understand VGG as a successor of the AlexNet but it was created by a different group named as Visual Geometry Group at Oxford's and hence the name VGG, It carries and uses some ideas from it is predecessors and improves on them and uses deep Convolutional neural layers to improve accuracy.

VGG network is the idea of much deeper networks and with much smaller filters. VGGNet increased the number of layers from eight layers in AlexNet.

The important point to note here is that all the convolution kernels are of size 3x3 and maxpool kernels are of size 2x2 with a stride of two.

The idea behind having fixed size kernels is that all the variable size convolutional kernels used in Alexnet (11x11, 5x5, 3x3) can be replicated by making use of multiple 3x3 kernels as building blocks. The replication is in terms of the receptive field covered by the kernels.

Let's explore what VGG19 is and compare it with some of other versions of the VGG architecture and also see some useful and practical applications of the VGG architecture.

Before diving in and looking at what VGG19 Architecture is let's take a look at ImageNet and a basic knowledge of CNN.

- Convolutional Neural Network(CNN)

First of all let's explore what ImageNet is. It is an Image database consisting of 14,197,122 images organized according to the WordNet hierarchy. This is an initiative to help researchers, students and others in the field of image and vision research.

ImageNet also hosts contests from which one was ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) which challenged researchers around the world to come up with solutions that yields the lowest top-1 and top-5 error rates (top-5 error rate would be the percent of images where the correct label is not one of the model's five most likely labels). The competition gives out a 1,000 class training set of 1.2 million images, a validation set of 50 thousand images and a test set of 150 thousand images.
and here comes the VGG Architecture, in 2014 it out-shined other state of the art models and is still preferred for a lot of challenging problems.

Architecture:

- A fixed size of (224 x 224) RGB image was given as input to this network which means that the matrix was of shape (224,224,3).
- The only preprocessing that was done is that they subtracted the mean RGB value from each pixel, computed over the whole training set.
- Used kernels of (3 x 3) size with a stride size of 1 pixel, this enabled them to cover the whole notion of the image.
- Spatial padding was used to preserve the spatial resolution of the image.
- Max pooling was performed over 2 x 2 pixel windows with stride 2.
- This was followed by Rectified linear unit(ReLu) to introduce non-linearity to make the model classify better and to improve computational time as the previous models used tanh or sigmoid functions that proved much better than those.
- Implemented three fully connected layers from which first two were of size 4096 and after that a layer with 1000 channels for 1000-way *ILSVRC* classification and the final layer is a softmax function.

Uses of the VGG Neural Network:

- The main purpose for which the VGG net was designed was to win the *ILSVRC* but it has been used in many other ways.
- Used just as a good classification architecture for many other datasets and as the authors made the models available to the public they can be used as is or with modification for other similar tasks also.
- Transfer learning : can be used for facial recognition tasks also.

- Weights are easily available with other frameworks like keras so they can be tinkered with and used for as one wants.
- Content and style loss using VGG-19 network

(Kaushik, 2020)

Unfortunately, there are two major drawbacks with VGGNet:

- It is very slow in the training phase.
- The network architecture weights themselves are quite large (in terms of disk/bandwidth).

Due to its depth and number of fully-connected nodes, VGG is over 533MB for VGG16 and 574MB for VGG19. This makes deploying VGG a tiresome task. We still use VGG in many deep learning image classification problems; however, smaller network architectures are often more desirable (such as SqueezeNet, GoogLeNet, etc.).

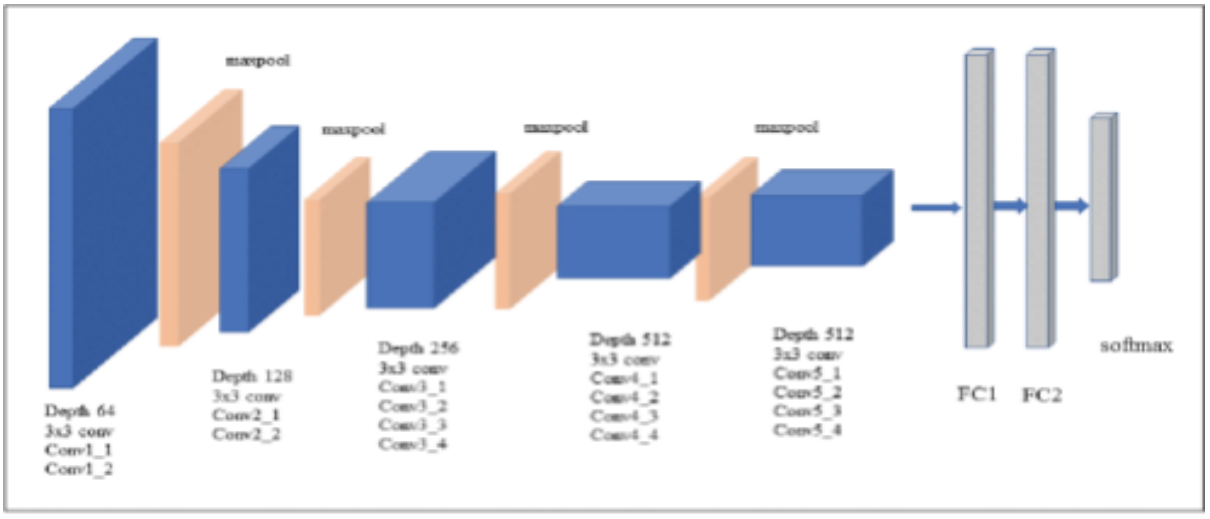


Figure 4.3 Vgg19 Architecture,
Source:https://saicharanars.medium.com/building-vgg19-with-keras-f516101c24cf

4.2.2 Resnet50

ResNet is a short name for Residual Network. As the name of the network indicates, the new terminology that this network introduces is residual learning.
What is the need for Residual Learning?
Deep convolutional neural networks have led to a series of breakthroughs for image classification. Many other visual recognition tasks have also greatly benefited from

very deep models. So, over the years there is a trend to go more deeper, to solve more complex tasks and to also improve classification accuracy. But, as we go deeper; the training of the neural network becomes difficult and also the accuracy starts saturating and then degrades also. Residual Learning tries to solve both these problems.

What is Residual Learning?

In general, in a deep convolutional neural network, several layers are stacked and are trained to the task at hand. The network learns several low, mid or high level features at the end of its layers. In residual learning, instead of trying to learn some features, we try to learn some residual. Residual can be simply understood as subtraction of features learned from input of that layer. ResNet does this using shortcut connections (directly connecting input of the nth layer to some (n+x)th layer. It has proved that training this form of networks is easier than training simple deep convolutional neural networks and also the problem of degrading accuracy is resolved.

The resnet 50 architecture contains the following element:

- A convolution with a kernel size of 7 x 7 and 64 different kernels all with a stride of size 2 giving us 1 layer.
- Next we have max pooling with also a stride size of 2.
- In the next convolution there is a 1 x 1,64 kernel following this a 3 x 3,64 kernel and at last a 1 x 1,256 kernel. These three layers are repeated in total 3 times so giving us 9 layers in this step.
- Next we have a kernel of 1 x 1,128 after that a kernel of 3 x 3,128 and at last a kernel of 1 x 1,512 this step was repeated 4 times so giving us 12 layers in this step.
- After that there is a kernel of 1 x 1,256 and two more kernels with 3 x 3,256 and 1 x 1,1024 and this is repeated 6 times giving us a total of 18 layers.
- And then again a 1 x 1,512 kernel with two more of 3 x 3,512 and 1 x 1,2048 and this was repeated 3 times giving us a total of 9 layers.
- After that we do an average pool and end it with a fully connected layer containing 1000 nodes and at the end a softmax function so this gives us 1 layer.

Uses of Resnet50:

- This architecture can be used on computer vision tasks such as image classification, object localisation, object detection.
- And this framework can also be applied to non computer vision tasks to give them the benefit of depth and to reduce the computational expense also.

Unlike traditional sequential network architectures such as AlexNet, OverFeat, and VGG, ResNet is instead a form of “exotic architecture” that relies on micro-architecture modules (also called “network-in-network architectures”). The term micro-architecture refers to the set of “building blocks” used to construct the network. A collection of micro-architecture building blocks (along with our standard CONV, POOL, etc. layers) leads to the macro-architecture (i.e,. the end network itself). First introduced by He et al. in their 2015 paper, Deep Residual Learning for Image Recognition, the ResNet architecture has become a seminal work, demonstrating that extremely deep networks can be trained using standard SGD (and a reasonable initialization function) through the use of residual modules. Further accuracy can be obtained by updating the residual module to use identity mappings, as demonstrated in their 2016 followup publication, Identity Mappings in Deep Residual Networks.

That said, keep in mind that the ResNet50 (as in 50 weight layers) implementation in the Keras core is based on the former 2015 paper. Even though ResNet is much deeper than VGG16 and VGG19, the model size is actually substantially smaller due to the usage of global average pooling rather than fully-connected layers — this reduces the model size down to 102 MB for Resnet 50.

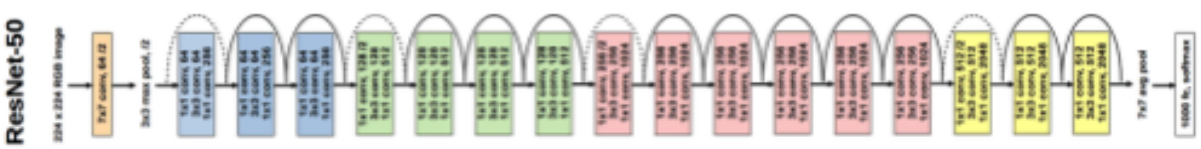


Figure 4.4 Resnet 50 Block Architecture
Source:https://www.researchgate.net/figure/VGG16-VGG19-Inception-V3-Xception-and-ResNet-50-architectures_fig1_330478807

4.2.3 Inception V3

InceptionV3 (Szegedy et al., 2016) is the third version of the popular Inception architecture (*Applications - InceptionV3*, n.d.). InceptionV3 was made popular due to its state-of-the-art results on the ILSVRC 2012 classification challenge and lower computational cost when compared to other widely used CNN architectures (Szegedy et al., 2016). The InceptionV3 block architecture is described in figure 4.5. The

concept behind an inception module is to perform a number of different operations and let the model decide (learn, during training) which output features are more important. This is opposed to explicitly creating layers only consisting of a 3 x 3 convolutional layer, for instance. Another notable aspect of this architecture is the use of the 1 x 1 convolutional layer, which is used to reduce the dimensionality of the features, saving on computation.

The goal of the inception module is to act as a "multi-level feature extractor" by computing 1×1, 3×3, and 5×5 convolutions within the same module of the network, giving us an enriched stack of features and informations as a result of the different filters applied to the image — the output of these filters are then stacked along the channel dimension and before being fed into the next layer in the network. The original incarnation of this architecture was called GoogLeNet, but subsequent manifestations have simply been called Inception vN where N refers to the version number put out by Google. The Inception V3 architecture included in the Keras core comes from the later publication by Szegedy et al., Rethinking the Inception Architecture for Computer Vision (2015) which proposes updates to the inception module to further boost ImageNet classification accuracy. The weights for Inception V3 are smaller than both VGG and ResNet, coming in at 96MB.

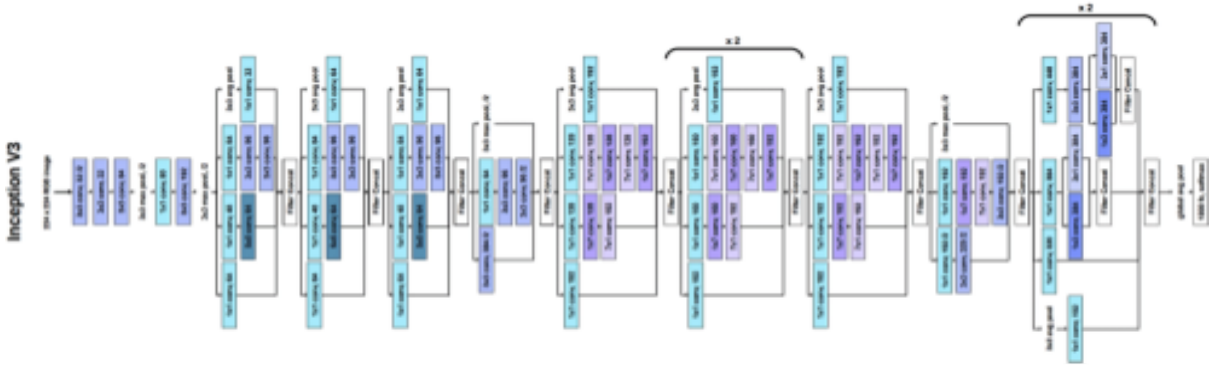


Figure 4.5 Inception V3 Block Architecture
Source:https://www.researchgate.net/figure/VGG16-VGG19-Inception-V3-Xception-and-ResNet-50-architectures_fig1_330478807

4.2.4 Xception

The basis of the Xception design is inspired by the Inception architecture (Chollet, 2017). The authors of (Chollet, 2017) argue that an Inception module performs similar to a traditional convolution and depth-wise separable convolution hybrid. With the success of depth-wise separable convolution in the mobileNet (Sandler et al., n.d.) architectures and the relative lightness compared to traditional convolution,

the authors of (Chollet, 2017) replaced all convolution layers with depth-wise separable convolutions. As seen in Figure 4.6, the Xception network appears to more closely resemble the ResNet (He et al., n.d.) network than the Inception (Szegedy et al., 2015), but because of the depth-wise separable layers the actual functionality is more of a hybrid between the two. Xception is much smaller than the behemoth InceptionResNet (Szegedy et al., n.d.) and is approximately the size of Inception V3 (Szegedy et al., 2015) and ResNet50 (He et al., n.d.). Benchmark performance on the ImageNet dataset Xception achieved a .945 accuracy compared to .941, and .933 to Inception V3 and ResNet-152 respectively (Chollet, 2017). While the Xception advancement seems only incremental, it does portray the understanding that different linear modules with residuals can operate similar to directed acyclic ones.

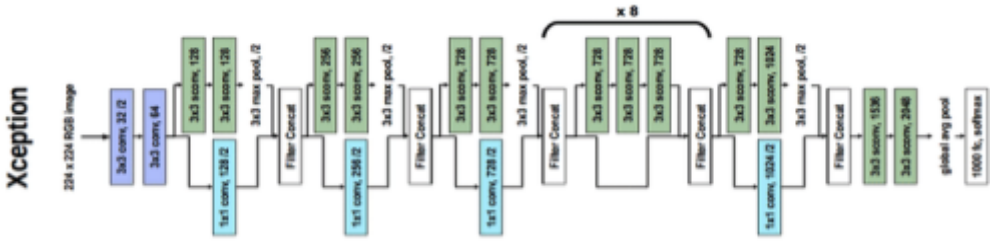


Figure 4.6 Xception Architecture
Source:https://www.researchgate.net/figure/VGG16-VGG19-Inception-V3-Xception-and-ResNet-50-architectures_fig1_330478807

4.2.5 MobileNetV1

MobileNet is a more suitable architecture for mobile and embedded version applications where the computing power is lacking. In terms of training speed but also in terms of prediction in real time, MobileNet is a class of convolutional neural systems planned by analysts at Google in April 2017. A few things that make MobileNet awesome' insanely small, fast, remarkably accurate, easy to tune for resource with accuracy. MobileNet covers less amount of space and can be trained very quickly compared to CNN. In addition, it is simple to tune and provides more accurate classification results compared to CNN. The image size should be greater than 32x 32 pixels for MobileNet otherwise it will be false. In this work 224, 224, 3 image sizes of RBG were used. MobileNet architecture uses depth wise separable convolutions that significantly decrease the number of parameters compared to normal network convolutions with the same network depth, the results in deep neural networks with lightweight. In the common convolution is replaced by

depthwise convolution followed by in point-wise convolution called as a convolution separable from the depthwise.

4.2.5.1 Convolutional Decomposition

MobileNet uses the idea of factorized convolution for reference and divides ordinary convolution operations into two parts: Depthwise convolution and point wise convolution. In depthwise convolution, each convolution core filter convolutes only for a specific input channel, as shown in the following figure, where M is the number of input channels and DK is the size of the convolution core.

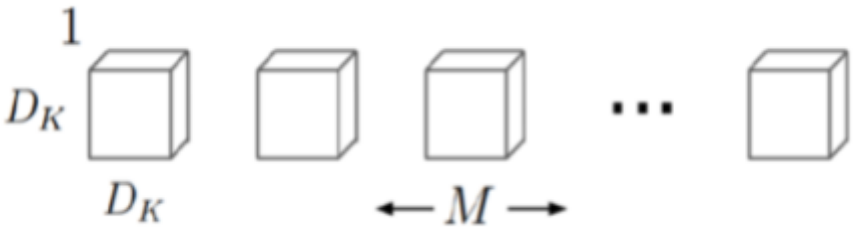


Figure 4.8 Depthwise Convolutional Filters
Source:https://www.slideshare.net/ssuser333356/mobilenetv1-v2-slide

The computational complexity of the Depthwise convolution formula is: DK x DK x M x DF x DF . Where DF is the size of the feature map of convolution layer output: In pointwise convolution, the multi-channel output of depthwise convolution layer is combined with a convolution core of 1×1 size. As shown below, N is the number of output channels. For convolution cores of 3×3 size, depthwise separable convolution can theoretically increase the efficiency by about 8-9 times. The formula is:

$$\frac{Dk * Dk * M * Df * Df + M * N * Df * Df}{Dk * Dk * M * N * Df * Df} = \frac{1}{N} + \frac{1}{Dk^2}$$

A standard convolutional layer of input a DF ×DF × M feature map F. which produces a DF × DF × N feature map G where DF is the spatial width and height of a square input feature map1, where M is the number of input channels (input depth), and DG is the spatial width and height of a square output feature map and N is the number of output channel output depth.

MobileNet CNN Architecture

In current work, the input images to the MobileNet(CNN) that passes from different layers.

Input layer: MobileNet has the ability to use multiple input layer sizes consisting of different width factors. The input sizes of the images in MobileNet range from 224x224 pixels.

Zero padding layer: non zero boundary conditions are used for most image recognition algorithms however MobileNet uses symmetric padding layers for modelling temporal data that should not infringe on the temporal order. The padding layer is used for maintaining the original data of an image.

Conv2D layer: This implies that a convolution process is in three dimensions but the movement of the filters in an image occurs in 2 dimensions across the image. The Conv2D layer uses a convolved layer to create a Tensor Flow of outputs for the convolution Kernel. Keyword arguments should be inputted when the layer is used as a first layer in a model. A filter size of 3x3 was used for convolutional layers.

Batch Normalize layer: this layer is used as a part of the architecture for normalization of every training mini batch of the model. It enables the use of a higher learning rate. This can sometimes eliminate the process of dropout as it can act as a regularization.

ReLu layer: after batch normalization, the ReLu layer follows. The ReLu activation layer for MobileNet comprises a ReLu function. The ReLu function is non-linear and it makes computation efficient for fast network convergence. ReLu prevents activation from getting cumbersome.the work used filter size 2x2 for ReLu layer.

Depth wise Cov2D: the depthwise conv2D allows the first step of the model to be performed in a depthwise spatial resolve in the convolution upon each input channel independently. In depth wise it uses filter size 1x1.

Global-average: this collects the average of pools from each preceding convolutional layers to prevent overfitting found in fully connected layers. This is also implemented to reduce model size and increase the prediction speed of a model. Dropout: this is a method used in deep learning for regularization. To also avoid overfitting in large networks the drop out technique ignores randomly selected neurons in a model during the training period.

Dense: this layer converts the features in an image into a single prediction for each image. It does not require the use of an activation function due to the raw prediction value used for prediction. Finally, images are classified into different classes. The structure of MobileNet is built on depth-wise convolutions as mentioned in the

previous section except for the first layer that is a full convolution. By defining the network in such simple terms it is easy to explore network topologies to find a good network. All layers are followed by a nonlinearity of batchnorm and ReLU except for the final fully connected layer that has no nonlinearity and feeds for classification into a softmax layer. Figure 4.4 contrasts a layer with regular convolutions, batchnorm and nonlinearity of ReLU with a factorized layer with convolution, 1 x 1 point conversion as well as batchnorm and ReLU after each convolution layer. Counting the convolutions in depth and point as separate layers, MobileNet has 93 layers.

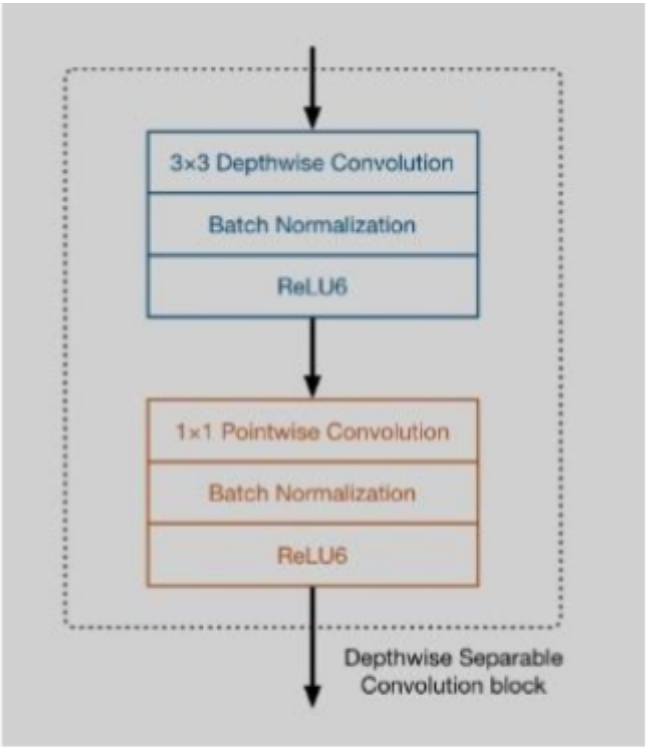


Figure 4.7 MobileNet V1 Architecture
Source:https://www.slideshare.net/ssuser333356/mobilenetv1-v2-slide

4.2.6 Reasons Behind the Selection of Pre-Trained Deep Neural Networks.

As I explained above, the reason we resorted to pre-trained networks is mainly the small amount of data available for such problems.The reason why we chose 5 networks to train instead of 1, is that each of the above networks has a different approach to the problem of classification. Vgg19 collects data hierarchically and proceeds with the classification, Resnet 50 uses blocks and shortcuts to reduce the

classification error, the Inception V3 uses parallel convolutions which it concatenates at the end of each layer producing an enriched information, the Xception adopts and combines techniques from both the inception and the resnet while the Mobilenet V1 is the lightest thanks to depth wise separable convolutions.All the above are based on different philosophies and using different architectures of connection of the different layers, attempting to achieve better characterization of the features.

Chapter 5
Classification Of Skin Lesions:Experiments and Results

5.1 Dataset Selection, Acquisition, And Preparation

In this thesis, we make use of 2 datasets that we will analyse further in chapters related to nevus/non nevus classification purpose. Subsequently we cite these datasets.

5.1.1 Related Datasets/Challenges (Isic) and Benchmarks

There are relatively few datasets in the general field of dermatology and even fewer datasets that include segmentation masks for benign and malignant lesions created by dermatologists. Additionally, many of the datasets can be challenging to acquire and often are not available publicly. These challenges make performing meaningful (comparable as a benchmark), reproducible, research unnecessarily difficult.
A few of the well known datasets are as follows: Dermofit Image Library, a dataset containing 1,300 high quality skin lesion images labeled into 10 different classes (*Edinburgh Research and Innovation. Dermofit Image Library*, n.d.), Dermnet, a skin disease atlas containing 23,000+ skin images across many diseases (*Dermnet - Skin Disease Atlas*, n.d.), PH2 (Mendonca et al., 2013), containing 200 dermoscopic images, and ISIC (International Skin Imaging Collaboration) archive which contains 13,000 skin lesion images (*ISIC Archive - International Skin Imaging Collaboration: Melanoma Project*, 2016).
In 2016, the International Symposium on Biomedical Imaging (ISBI) (*IEEE International Symposium on Biomedical Imaging*, 2016) released the first annual challenge dataset for "Skin lesion analysis towards melanoma detection using photos from the ISIC archive (*ISIC. ISBI 2016: Skin lesion analysis towards melanoma detection*, 2016). This provided a public dataset and public "leaderboard" which provided a means to benchmarking results. Both the ISIC and ISBI 2016 dataset are further discussed in Section 3.1. An ISBI 2017 challenge has also been released (*ISIC. ISBI 2017: Skin lesion analysis towards melanoma detection*, 2017), and a 2018 challenge planned for release (*ISIC. ISIC skin image analysis workshop and challenge @ miccai 2018*, 2017).

5.1.2 Kaggle for deep learning

Kaggle “Dermoscopy_Images” dataset details:

- 2 Classes (Nevus,Malignant)
- 6000 images
- Split in 2400 test nevus images
- 2400 malignant lesion test images
- 75 nevus validation images
- 75 malignant validation images
- 525 nevus test images
- 525 malignant test images

The dataset is a subset of the isic dataset from isic archive that contains various skin lesion images.(Tascon, 2019). All images were unaltered beyond basic cropping and/or resizing to the specified input size. The dataset was split as such that no images overlapped between training, validation, and test sets. At the time of development, only these specific datasets were accessible and publicly available. Datasets with nevus/non nevus images and segmentation masks respectively. Our kaggle dataset images are dermoscopic images which contain the lesion areas and a part of the skin around them and also in some of them there is hair.

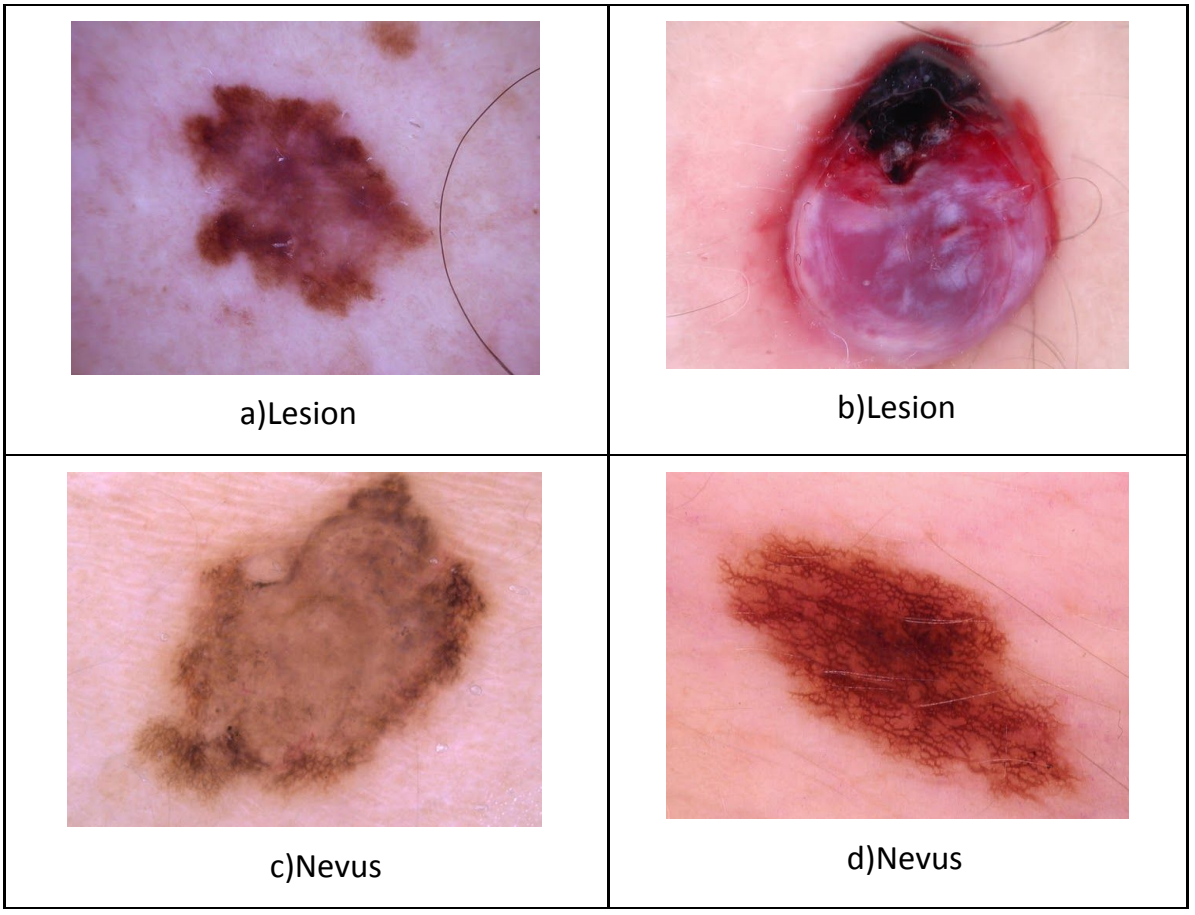


Figure 5.1 Images From Kaggle Dataset

5.2 Data Preprocessing:Image Augmentation

Based on these 2 datasets, we will attempt to test and evaluate the ability of the network structures, to extract appropriate features in order to characterise the lesions as nevus or non nevus.

We used RGB images as input, in order to take advantage of color information.All images were subjected to basic cropping and/or resizing to the specified input size. The dataset was split, such that no images overlapped between training, validation, and test sets. Metrics of interest were produced on a balanced (benign/- malignant) test set. It is worth nothing that the balance of the evenly balanced split test 22 used for evaluation, does not accurately represent the balance that a classifier may encounter when deployed in a clinical setting. It is almost certainly the case that in a clinical setting, the balance will be shifted considerably in favor of benign lesions. At the time of development, the Kaggle "Dermoscopy_Images" was the only readily accessible, publicly available, dataset with included benchmarks and segmentation masks.In order to have a more complete dataset, actions such as resizing and cropping were necessary.

5.3 Proposed Methodology and Experiments

As mentioned before, the aim of this study is to create an accurate classifier that distinguishes accurately the healthy nevus and the malignant lesions. Since the problem is primary medical, we have to choose a classifier that produces the minimum error in order to avoid wrong estimates about the lesion. For instance, let's consider a model that estimates 60% of malignant nevus as healthy. This scenario is really dangerous and must be avoided. Developing the most powerful model is a high demanding process and it needs time as well as computer resources (especially for deep models). In this study, we compare a variety of architectures and models (pretrained and non-pretrained) among a great number that have been published by researchers in the computer vision field. In the following sections, all models used with their parameters are given. For each model we give the confusion matrix, F1-score, Precision, Recall and network training process. Our input image shape is 224x224x3 and our output is a possibility between 0 and 1, with 0 for healthy nevus and 1 for other skin lesions. In all of our pretrained networks we chose to freeze the layers except the last 5 (including the classifiers) and the batch normalization layers. This happens due to the fact that we apply fine tuning to the networks, particularly

for our problem.The problem is that our dataset is different from the datasets of ImageNet, which is the knowledge database for our pretrained models. Therefore, we unfreeze the last 5 layers, where high order features i.e object parts and classes are extracted, and train them into our new dataset.Also, as a crucial “metric” in order to estimate the reliability of our solutions, there is a term called dangerous diagnoses. Dangerous diagnoses are the top right cases on the confusion matrix, where a non nevus (and possible cancerous lesion) is diagnosed as a normal nevus. They are considered crucial because it is a real life problem, which is why we also used the f1 score, as we mentioned before.During this thesis, we unfreeze the last five layers and all the batch normalization layers in order to help improve network performance. As in prior networks, the output is a probability and if the probability is greater than 0.5 , then the lesion is considered dangerous and must be examined further from a doctor. In the other case, the lesion is considered nevus.The hyperparameters of the networks are given in the following table.

- Max Epochs : 150
- Pooling layers : Max Pooling
- Learning Rate : 0.001
- Optimizer : Adam
- Dropout rate (used only in NN layers) : 0.4
- Loss function: Binary cross-entropy loss
- ReduceLROnPlateau : monitor loss on validation set.
- Early Stopping : monitor accuracy on training set.

Shallow net:

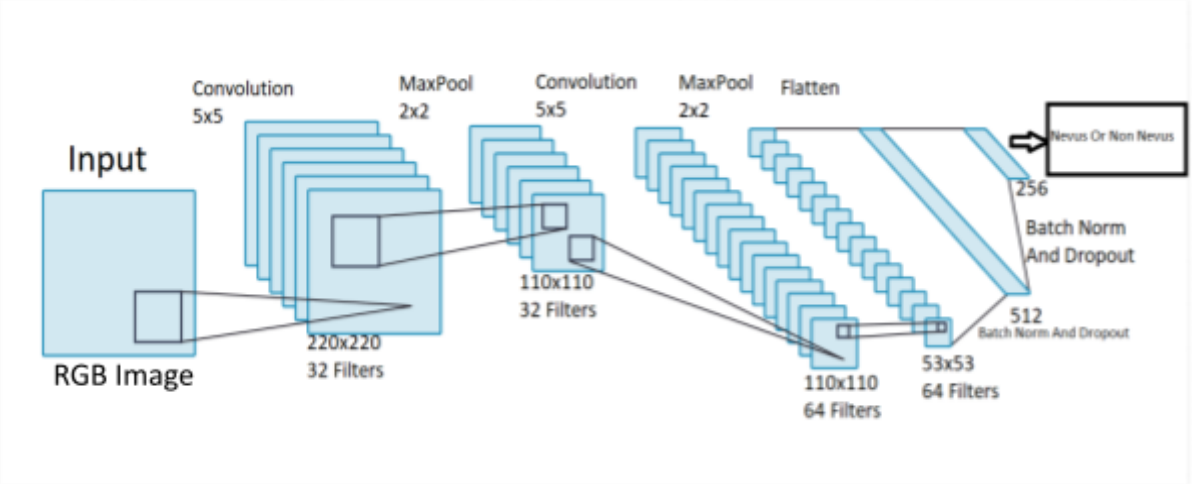


Figure 5.2 Basic Net

Basic net (BN) is the most simple architecture of a CNN. Firstly, we want the bare minimum model in terms of training time and size. The smaller-sized models can be loaded in any physical device (i.e smartphone) and produces real- time results. On the other hand, DNN requires powerful hardware (mostly GPU) and is extremely sizable (~ 1GB). However, DNN needs more training time than shallows, and is suffering from gradient vanishing or exploding. The BN format shares a lot of similarities with VGG and AlexNet. BN follows the VGG pattern: at the early hidden layers, the spatial resolution of the intermediate images are high and the channel number is low. As we traverse the model towards the final layers, this trend is changed and the spectral (number of channels) resolution tends to dominate. BN is structured with two convolutional and max pooling layers, as it is depicted in fig. 6.1. The final flatten layers are simple NN that are the classifiers of the skin images. Furthermore, BN takes as input an RGB (224 x 224) dermoscopic image and produces as output the probability of being non-nevus lesion. In that case, if the probability is greater than 0.5 then the lesion is considered dangerous and must be examined further from a doctor. In the other case, the lesion is considered nevus.

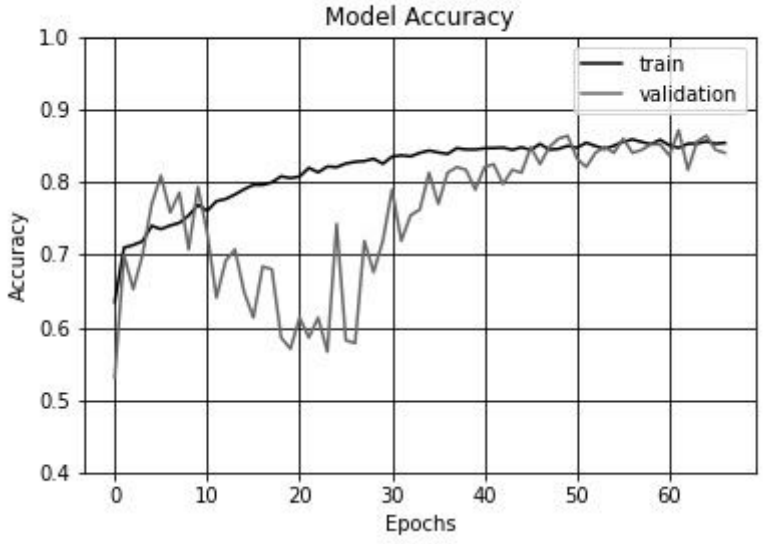


Figure 5.3 Basic Net training process

Actual/Predicted	Non Nevus	Nevus
Non Nevus	518	82
Nevus	125	475

Table 5.1 Basic Net: Confusion Matrix

	Precision	Recall	F1 Score
Non Nevus	0.81	0.86	0.83
Nevus	0.85	0.79	0.82

Table 5.2 Basic Net: Precision, Recall, F1-score

Figure 5.2 depicts how the model accuracy is changed during the training epochs. Basic net model converges after 50 epochs with approximately 82 % accuracy and 82 dangerous diagnoses out of 1200 images.

Parameters Table

Total Parameters	20,421,697
Trainable Parameters	20,420,161
Non-Trainable Parameters	1,536

Table 5.3 Parameters Table

Vgg19:

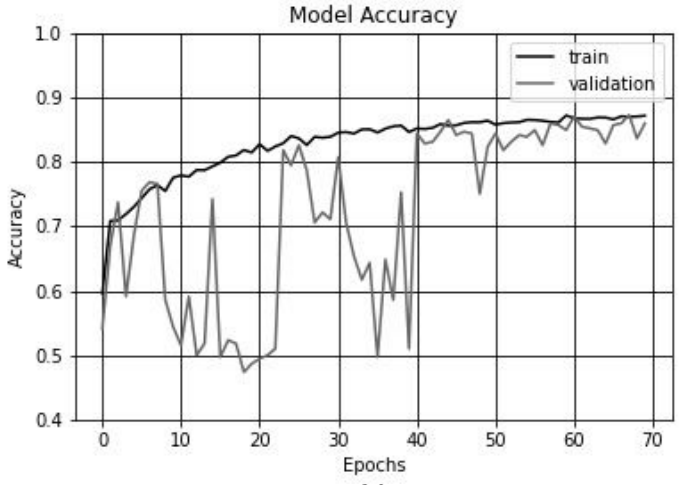


Figure 5.4 Vgg19 Training Process

Vgg19, as we mentioned before, is a pre-trained neural network from ImageNet, that takes an RGB image 224x224 pixels and gathers features in a hierarchical way.
It begins from edges to texture, to object parts and finally object classes and during the process we see the spatial domain to decrease while the channel domain increases.
Figure 5.3 depicts how the model accuracy is changed during the training epochs. The Vgg19 model converges after 60 epochs with approximately 85 % accuracy. As we see, Vgg19 has better accuracy than Shallow Net because it may gathers hierarchically the features but it also has more dangerous diagnoses (111 cases out of 1200) and as we mentioned, in real life problems F1 Score ,which relies more at false positives cases in the confusion matrix, is considered a more suitable metric.

Actual/Predicted	Non Nevus	Nevus
Non Nevus	489	111
Nevus	72	528

Table 5.4 Vgg19:Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.87	0.81	0.84
Nevus	0.83	0.88	0.85

Table 5.5 Vgg19 : Precision,Recall,F1 Score

Parameters Table

Total Parameters	20,421,697
Trainable Parameters	20,420,161
Non-Trainable Parameters	1,536

Table 5.6 Vgg19 : Parameters Table

Resnet 50:

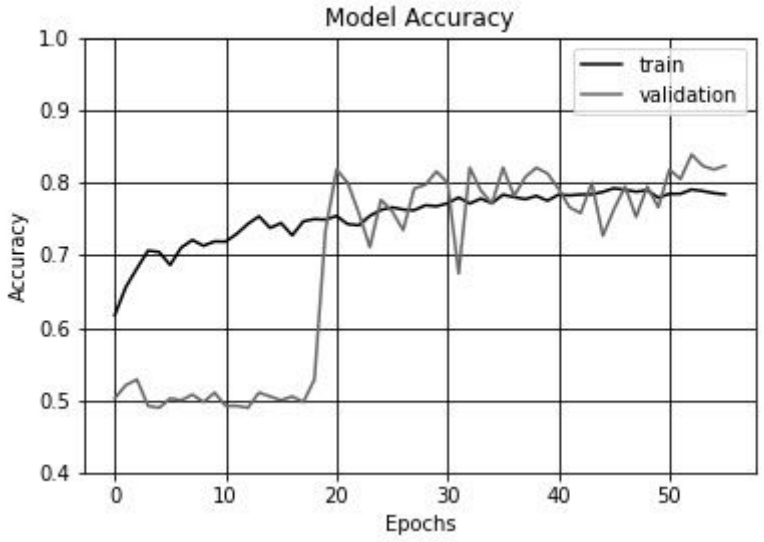


Figure 5.5 Resnet50 Training Process

Resnet50 is a pre-trained neural network that introduces the concept of micro architecture models or building blocks in order to create a neural network and it takes as input an RGB image 224x224 pixels.

A series of microarchitecture modules leads us to the macroarchitecture model, in other words the neural network.It takes the module output from the first 2 convolutional layers and adds it to the initial input in order to avoid gradient vanishing.The successive paths that add each module output to the initial input of the same module are called shortcuts.Although resnet 50 is a much deeper network than vgg19, it is also much smaller because instead of using fully connected layers, it uses average pooling layers which its size is at most 102 mb.Figure 5.4 depicts how the model accuracy is changed during the training epochs. The Resnet 50 model reaches about 80 % accuracy after 60 epochs. Resnet 50 lacks accuracy compared to Vgg19, but it has better score in dangerous diagnoses.Due to this structure, the gradient vanishing is avoided.

Actual/Predicted	Non Nevus	Nevus
Non Nevus	537	63

Nevus	176	424

Table 5.7 Renset 50 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.75	0.90	0.82
Nevus	0.87	0.71	0.78

Table 5.8 Resnet 50 Precision,Recall,F1 Score

Parameters Table

Total Parameters	25,636,712
Trainable Parameters	25,583,592
Non-Trainable Parameters	53,120

Table 5.9 Resnet 50 Parameters Table

Inception V3:

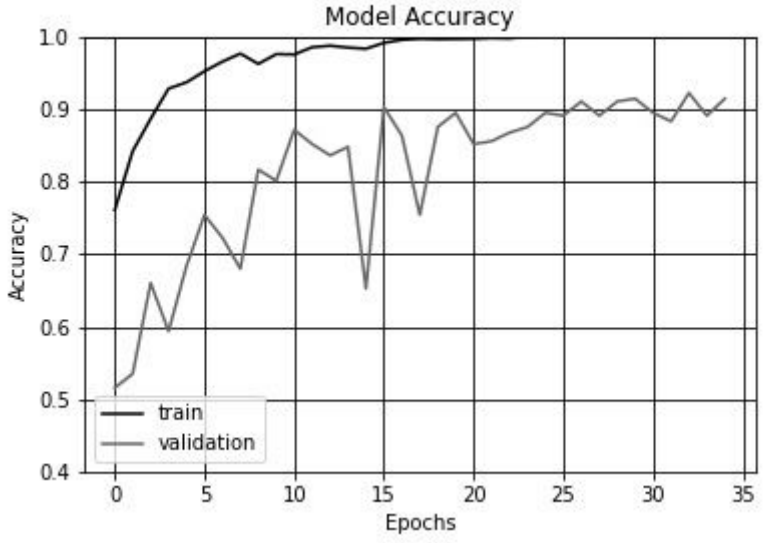


Figure 5.6 Inception V3 Training Process

Inception V3 is a pre-trained neural network which instead of having 1 convolutional layer, it has modules with different convolutional and pooling layers in parallel, whose outputs concatenate on the channel domain and create the final output of the module.

That way, we have a stack with the features of the corresponding convolutional layers.

It is an enriched stack of information from different filters over the same image. Using this technique, we achieve higher accuracy than the previous solutions, with much less dangerous diagnoses(cases where a non nevus lesion is considered as nevus).

Figure 5.5 depicts how the model accuracy is changed during the training epochs. The Inception V3 model stabilizes at 91% accuracy, after 25 epochs. Inception V3 excels in relation to resnet 50, because it has parallel inputs in every block with convolution and pooling layers and the results are concatenating in the output. In that way, we have a more enriched information as output and as a result we have better estimations.

Actual/Predicted	Non Nevus	Nevus
Non Nevus	567	33
Nevus	75	525

Table 5.10 Inception V3 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.88	0.94	0.91
Nevus	0.94	0.88	0.91

Table 5.11 Inception V3 Precision,Recall,F1 Score

Parameters Table

Total Parameters	22,986,529

Trainable Parameters	22,950,561
Non-Trainable Parameters	35,968

Table 5.12 Inception V3 Parameters Table

Xception:

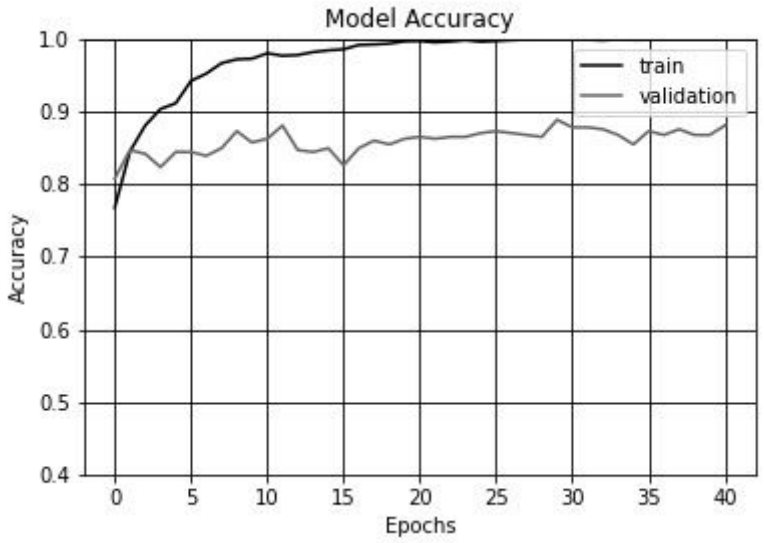


Figure 5.7 Xception Training Process

Xception is a pre-trained neural network and an expansion of the Inception V3 network.It replaces the standard Inception modules with depthwise separable convolutions and it has the smallest weight serialization at only 91 mb.Figure 5.6 depicts how the model accuracy is changed during the training epochs. The Xception model approaches 88% accuracy after 40 epochs. Xception also has good accuracy and less dangerous diagnoses than shallow and vgg, almost the same as the resnet but worse than inception.

Actual/Predicted	Non Nevus	Nevus
Non Nevus	527	74
Nevus	86	514

Table 5.13 Xception Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.86	0.88	0.87
Nevus	0.87	0.86	0.87

Table 5.14 Xception Precision,Recall,F1 Score

Parameters Table

Total Parameters	22,910,480
Trainable Parameters	22,855,952
Non-Trainable Parameters	54,528

Table 5.15 Xception Parameters Table

MobileNetV1:

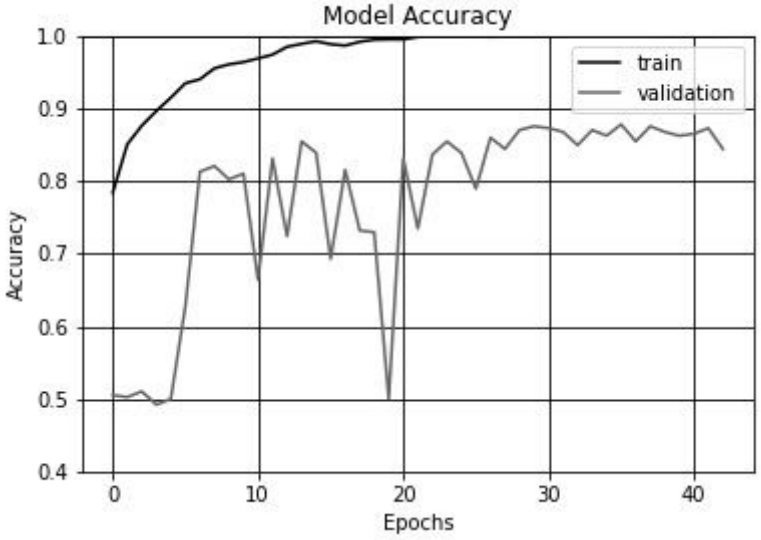


Figure 5.8 Mobilenet V1 Training Process

MobileNet V1 is a pre-trained neural network which uses, as the Xception, depthwise separable convolutions and it contains 28 convolutional layers and 1 fully connected layer that is followed by a softmax classifier. After every convolution, we have batch normalization layers and ReLu activation functions. MobileNet V1 outperforms networks, such as Vgg and GoogleNet, with much less parameters and is useful for mobile and embedded computer vision applications. Also, thanks to the separable convolutions, it has much smaller complexity. Figure 5.7 depicts how the model accuracy is changed during the training epochs. The MobileNet V1 model converges

after 40 epochs with approximately 88% accuracy and better dangerous diagnoses than Xception by 2.

Actual/Predicted	Non Nevus	Nevus
Non Nevus	528	72
Nevus	73	527

Table 5.16 MobileNet V1 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.88	0.88	0.88
Nevus	0.88	0.88	0.88

Table 5.17 MobileNet V1 Precision,Recall,F1 Score

Parameters Table

Total Parameters	3,888,321
Trainable Parameters	1,708,545
Non-Trainable Parameters	2,179,776

Table 5.18 MobileNet V1 Parameters Table

5.4 Comparison and Discussion of the results

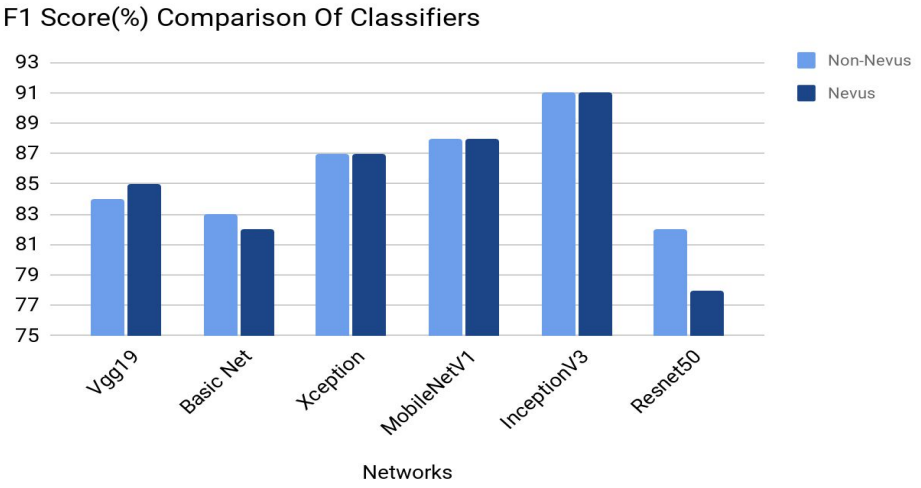


Figure 5.9 F1 Score Comparison

Discussion over the results:

The bar chart shows the F1 Scores of six networks, Shallow net, Vgg19, Resnet 50, Inception V3, Xception and MobileNet V1 of two different classes of images (nevus and non nevus lesions).

Vgg19 has better results than the Shallow network in F1 score (84% and 85% instead of 83% and 82% for non nevus and nevus respectively) and accuracy (85% instead of 82%) as a result of hierarchical feature extraction, but it has more dangerous diagnoses(111 against 82).

Dangerous diagnoses are the top cases on the confusion matrix, where a non nevus (and possible cancerous lesion) is diagnosed as a normal nevus.

Resnet 50 has worse F1 score (82% and 78% instead of 84% and 85%) and accuracy (80% against 85%) than Vgg19 but much less dangerous diagnoses(63). This is very important especially for biomedical apps, due to the fact that in each block it adds its output to the input of the block, which results in richer information and avoids overfitting and gradient vanishing.

Inception V3 has the best F1 score (91% for both classes)and accuracy (91%). The reason behind this, is that max pooling layers and convolutional layers are connected in parallel while their outputs are concatenated into one, giving us enriched information by the stacked features from the layers and the least dangerous diagnoses (33). Xception has a combination of batch normalization and depth wise

separable convolutions. It also has good F1 score (87% for both classes) and accuracy (88%) with less dangerous diagnoses (74) than Shallow net and Vgg19 but more than Inception V3 and Resnet 50. Finally, MobileNet V1 has the second best F1 score (88% for both classes) and accuracy (88%) and the third best score of dangerous diagnoses(72). It also makes use of batch normalization and Depth Wise Separable Convolutions.

Taking into consideration the above metrics, Inception V3 is the best option for each one of them.Also, if we want to take a better look at the network metrics we can observe the following table.

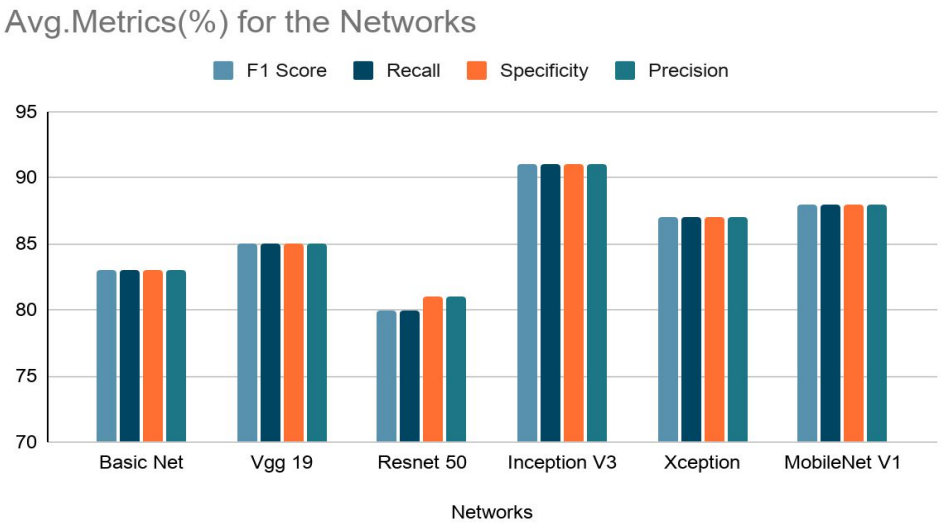


Figure 5.10 Average Network Metrics

We observe that all metrics are in the same or almost the same levels.

However, it is important to focus on metrics such as sensitivity(recall) and specificity.

Sensitivity of a classifier is the ratio between how much were correctly identified as positive to how much were actually positive. It is used in cases where classification of positives is a high priority.

Specificity of a classifier is the ratio between how much was correctly classified as negative to how much was actually negative. Specificity is used in cases where classification of negatives is a high priority.

Eg: Diagnosing for a health condition before treatment.
Once again, Inception V3 has the best results among the others.

5.5 10-Fold Cross Validation With Inception V3

Since we have come to the conclusion that the Inception V3 network has the best metrics for a single run, it is now time to check how this model can reliably generalize unknown data. In order to check this, we will use k-fold cross validation.
K-fold cross validation is a popular method of cross validation which shuffles the data and splits them into k number of folds (groups). In general, K-fold validation is performed by taking one group as the test data set, and the other k-1 groups as the training data, fitting and evaluating a model, and recording the chosen score. This process is then repeated with each fold (group) as the test data and all scores are averaged to obtain a more comprehensive model validation score.
In our case we leverage 10-fold cross validation in order to be representative of the model and small enough to be computed in a reasonable amount of time. Depending on the dataset size, different k values can sometimes be experimented with. As a general rule, as k increases, bias decreases and variance increases.
Following the above, we quote the metrics of the model for the 10-fold cross validation method.

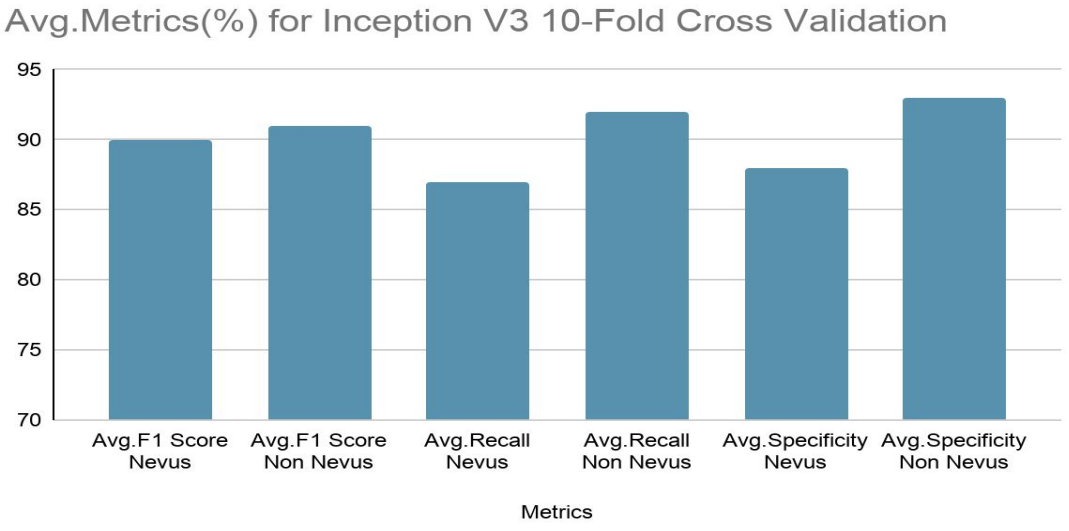


Figure 5.11 Cross Validation Avg.Metrics

Inception V3	F1 Score	Recall	Specificity
Non Nevus	0.91	0.92	0.93
Nevus	0.90	0.87	0.88

Table 5.19 Cross Validation Avg.Metrics Table

As can be seen the metrics are slightly downwards but at the same levels as the single run of the model.
In any case we have a more complete and better picture now of how well our model is performing in a "real world" simulation.

Chapter 6
Skin Lesion Segmentation and Experiments

Up to this point, we have seen the results in dermoscopic images without further processing except for the standard image augmentation.
Therefore, our thought is to see if we can help our networks to improve their metrics by feeding them segmented images, where the skin lesion has been isolated.
So the subject of this chapter is the skin lesion isolation with segmentation of the region of interest. Afterwards, there is a second round of experiments, as in chapter 6, in order to find out if this approach is efficient.

6.1 State of the art

Akram et al., 2020 utilize recent deep models for feature extraction, by taking advantage of transfer learning. Initially, the dermoscopic images are segmented and the lesion region is extracted, which is later subjected to retrain the selected deep models to generate fused feature vectors. In the second phase, a framework for most discriminant feature selection and dimensionality reduction is proposed,
entropy-controlled neighborhood component analysis (ECNCA). This hierarchical framework optimizes fused features by selecting the principle components and extricating the redundant and irrelevant data. The effectiveness of their design is validated on four benchmark dermoscopic datasets; PH2, ISIC MSK, ISIC UDA, and ISBI-2017. To authenticate the proposed method, a fair comparison with the existing techniques is also provided. The simulation results clearly show that the proposed design is accurate enough to categorize the skin lesion with 98.8%, 99.2% and 97.1% and 95.9% accuracy with the selected classifiers on all four datasets and by utilizing less than 3% features.
Tabatabaie et al., 2009 after preprocessing and segmentation of the images, features that describe the texture of lesions and show high discriminative characteristics, are extracted using ICA. Afterwards, these features, along with the color features of the lesions, are used to construct a classification module based on support vector machines for the recognition of malignant melanoma vs. benign nevus. Results: Experimental results showed that combining melanoma and nevus color features with proposed ICAbased texture features led to a classification accuracy of 88.7%.
Khakabi et al., 2015 apply spatial and color features in order to model the lesion growth pattern. The decomposition is done by repeatedly clustering pixels into dark

and light sub-clusters. A novel tree structure based representation of the lesion growth pattern is constructed by matching every pixel sub-cluster with a node in the tree structure. This model provides a powerful framework to extract features and to train models for lesion segmentation. The model employed allows features to be extracted at multiple layers of the tree structure, enabling a more descriptive feature set. Additionally, there is no need for preprocessing such as color calibration or artifact disocclusion.

Preliminary features (mean over RGB color channels) are extracted for every pixel over four layers of the growth pattern model and are used in association with radial distance as a spatial feature to segment the lesion. The resulting per pixel feature vectors of length 13 are used in a supervised learning model for estimating parameters and segmenting the lesion. A dataset containing 116 challenging images from dermoscopic atlases is used to validate the method via a 10-fold cross validation procedure. Results of segmentation are compared with six other skin lesion segmentation methods. Their method performs competitively with another method. They achieve a per-pixel sensitivity/specificity of 0.890 and 0.901 respectively.

Another approach in the section of skin lesion segmentation, is the Unet [(Lopez et al., 2017), (Ronneberger et al., 2015)].

The recent emergence of machine learning and deep learning methods for medical image analysis, has enabled the development of intelligent medical imaging-based diagnosis systems that can assist physicians in making better decisions regarding a patient's health. In particular, skin imaging is a field where these new methods can be applied with a high rate of success.

Lopez's thesis focuses on the problem of automatic skin lesion detection, particularly on melanoma detection, by applying semantic segmentation and classification from dermoscopic images using a deep learning based approach. For the first problem, a U-Net convolutional neural network architecture is applied for an accurate extraction of the lesion region. For the second problem, the current model performs a binary classification (benign versus malignant) that can be used for early melanoma detection. The model is general enough to be extended to multi-class skin lesion classification. The proposed solution is built around the VGG-Net ConvNet architecture and uses the transfer learning paradigm. Finally, his work performs a comparative evaluation of classification alone (using the entire image) against a combination of the two approaches (segmentation followed by classification) in order to assess which of them achieves better classification results.

In the end, we studied the work of (Iglovikov & Shvets, 2018) about the Ternaus net.

Classical U-Net architectures composed of encoders and decoders are very popular for segmentation of medical images, satellite images etc. Typically, a neural network initialized with weights from a network pre-trained on a large data set, like ImageNet shows better performance than those trained from scratch on a small dataset. In some practical applications, particularly in medicine and traffic safety, the accuracy of the models is of utmost importance. In their paper, they demonstrate how the U-Net type architecture can be improved by the use of the pretrained encoder. They compare three weight initialization schemes: LeCun uniform, the encoder with weights from VGG11 and a full network trained on the Carvana dataset.

6.2 Kaggle Dataset For Semantic Segmentation

Kaggle Skin Lesion images and their segmented ground truths.
This dataset contains :

- 2750 images, and their segmented grayscale masks.
- Two Classes.
- Class X is for the rgb skin lesion images and Class Y is for the grayscale masks/segmented ground truths.
- 2000 of them for each class are the training dataset, 150 for each class is the validation set and 600 images of the test set for each class respectively.

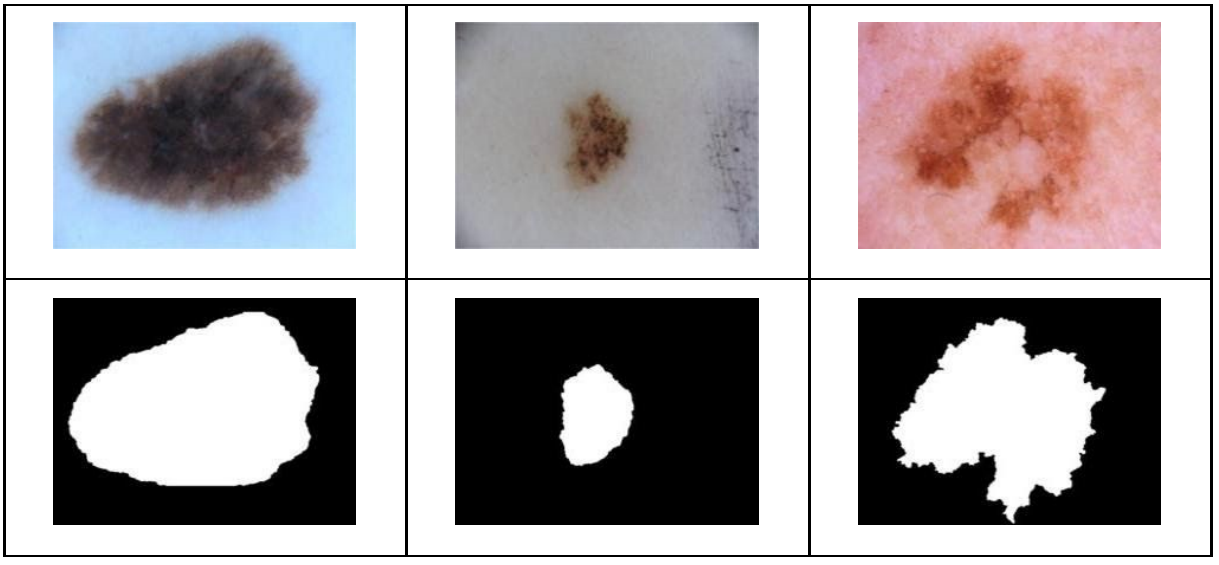


Figure 6.1 The two classes of the dataset for semantic segmentation

6.3 Semantic Segmentation (ML Techniques)

6.3.1 Introduction

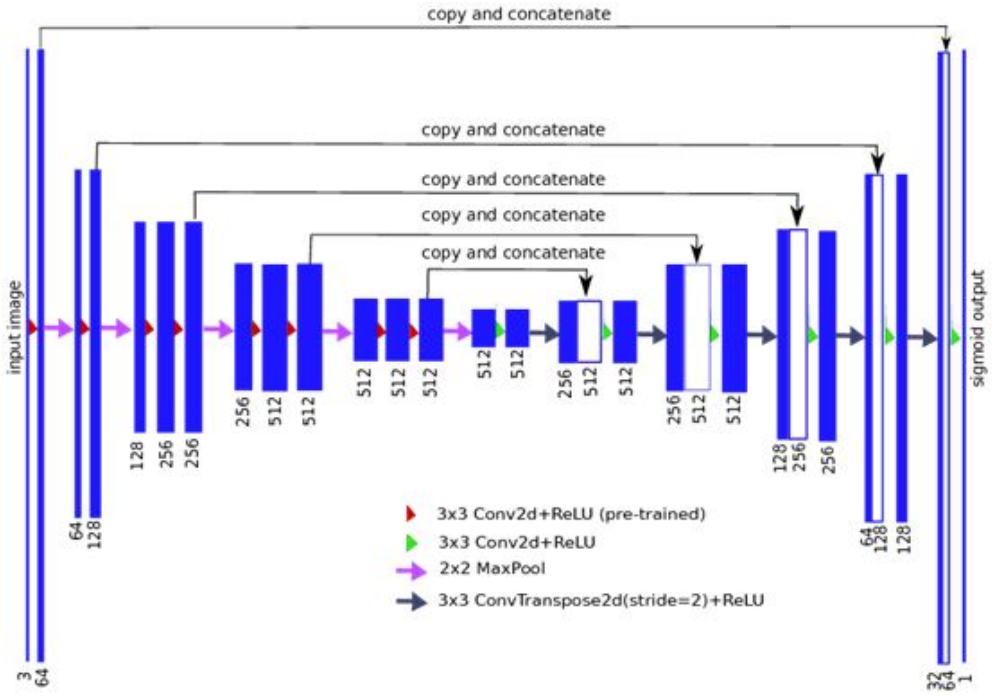


Figure 6.2 Ternaus Net Architecture

By using end-to-end methods, we had some very good results such as in the Inception V3 model. But would the models improve if we feed them images in which we have isolated only the area of the nevus? The Ternaus network (Iglovikov & Shvets, 2018) is a variant of Unet [(Ronneberger et al., 2015), (Lopez, 2017)] which uses some pretrained layers of Vgg 16 in order to do semantic segmentation in an image and in our case to isolate the nevus from the skin and hair as much as possible. We chose this approach, because Unet is a convolutional network architecture for fast and precise segmentation of images. Up to now, it has outperformed the prior best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. It has won the Grand Challenge for Computer-Automated Detection of Caries in Bitewing Radiography at ISBI 2015, and it has won the Cell Tracking Challenge at ISBI 2015 on the two most challenging transmitted light microscopy categories (Phase contrast and DIC microscopy) by a large margin.

Also, Ternaus net was a part of the winning solution (1st out of 735) in the Kaggle: Carvana Image Masking Challenge so it was seen as the best option to follow, in order to solve our problem of skin lesion segmentation.

6.3.2 Semantic Segmentation on medical images

What is image segmentation:

It is known that an image is nothing but a collection of pixels. Image segmentation is the process of classifying each pixel in an image belonging to a certain class and hence can be thought of as a classification problem per pixel. There are two types of segmentation techniques

1. Semantic segmentation :- Semantic segmentation is the process of classifying each pixel belonging to a particular label. It doesn't different across different instances of the same object. For example if there are 2 cats in an image, semantic segmentation gives same label to all the pixels of both cats
2. Instance segmentation :- Instance segmentation differs from semantic segmentation in the sense that it gives a unique label to every instance of a particular object in the image. As can be seen in the image above all 3 dogs are assigned different colours i.e different labels. With semantic segmentation all of them would have been assigned the same colour.

6.3.3 Implementation Details

This network consists of two stages. The first stage is the encoder path, in which we do downsampling of the image and the second stage is the decoder path, in which we upsampling the image in order to create the binary mask of the skin lesion. The name of the Unet network comes from the form of the network that has a “U” shape. We use at first the Vgg16 network layers because they are identical with the encoder stage layers. The network receives as input an RGB image and produces as output the binary mask of this image where the black area represents the skin and the white area represents the skin lesion surface. After we complete the training of the ternaus net, we feed the network with the kaggle "Dermoscopy_Images" dataset that we used at Chapter 6. We produce the binary masks of the training, testing and validation set and then we concatenate them with the original images, in order to have as a result an RGB image with the skin lesion area isolated. With this "new" dataset we repeat the experiments in order to find if the skin lesion segmentation improves the classification metrics. For the needs of skin lesion segmentation part,

we used the kaggle hashbanger dataset, which consists of X and Y image classes, for the RGB images and their grayscale masks, respectively where the Y class images will be compared with the masks that we will produce in order to extract the network metrics. Initially, our network receives as input an RGB image, in which we want to do semantic segmentation, i.e. to categorize and display the pixels of our image either as black, i.e. background or skin lesion regarding the region of interest. Our image is subject to a series of convolutional layers with same padding that increase the channel domain and then pooling layers which reduce the spatial domain. Also, in each step the last image is kept so that they become concatenated with the symmetrical feature map of the decoder, which I will explain below. This spatial degradation and channel increase is done as in Vgg16 in order to learn the really important features of the image and to export their basic map. At the decoder stage however, we will see that in order for these features to become visible to the human eye, we will have to gradually increase the spatial domain this time and reduce the channels. As for the encoder path instead of the original Unet code I used the corresponding and pre-trained layers vgg 16 from Imagenet datasets in the extraction of features and I then concatenated them in the corresponding and symmetrical feature maps of the decoder. So after I get to the last layer of the encoder, we get to the point where the decoder's path starts. This is thanks to two basic functions, conv 2d transpose and concatenate. Concatenate is when different feature maps are stacked between one another in the channel domain. But before we move on to the decoders path, the meaning of transpose convolution must be understood.

Transpose convolution is practically the deconvolution. In order to make this clear we can say that the result of convolution of an image with a filter is displayed as the projection of the original panel if we illuminated it with a lens. The clearly smaller shadow that would result is the outcome of convolution and the size of the filter is our visual field through which we see this projection. Therefore, what we do in conv2d Transpose is, in an abstract sense, that we illuminate with the lens on the other side, from the smallest object this time. The mathematical explanation is that for each value of the smallest object we create new arrays that are the result of multiplying the filter by the array elements and combining the resulting arrays by summing the overlapping elements according to the dimensions of the array we want to produce. In short, it is upsampling and convolution in one step. In the concatenate state we have the same spatial as the symmetric but double size of the channel's number. In order to get better placements in each step of the decoder, we concatenate the output of the transposed convolutional layers with the characteristic

map from the encoder of the same level to avoid gradient vanishing. After every concatenation, we apply again two consecutive regular convs so that the model can learn to assemble a more precise output. The final output is a grayscale image, the mask of the starter image. In the end it gets through the sigmoid function according to the graph of which a number between zero and one is obtained for each pixel. Each pixel less than 0.5 to the left of the axis becomes black and the rest are white. Through this process, the table shows the original image results masks.

6.3.4 Skin Lesion Semantic Segmentation (Ternaus Net) Experiment and Results

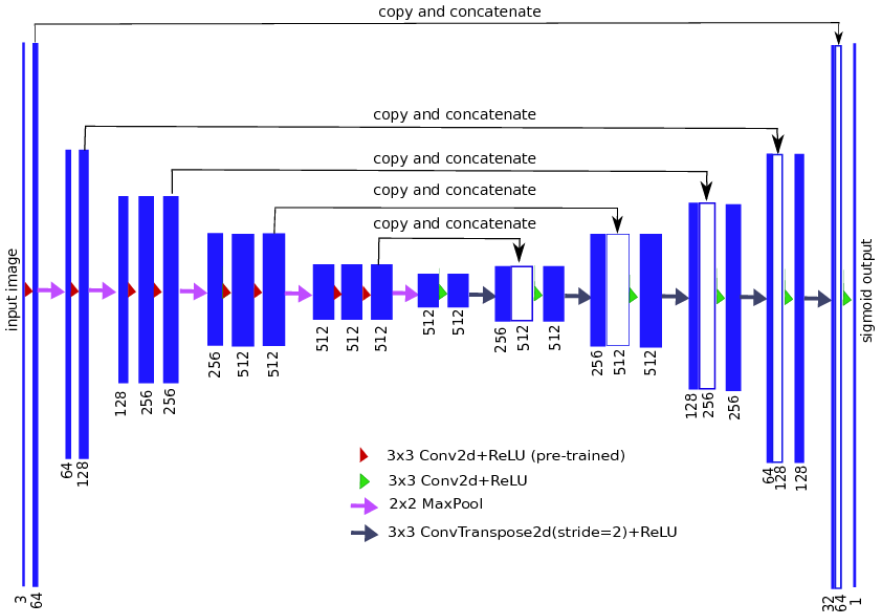


Figure 6.3 Ternaus Network

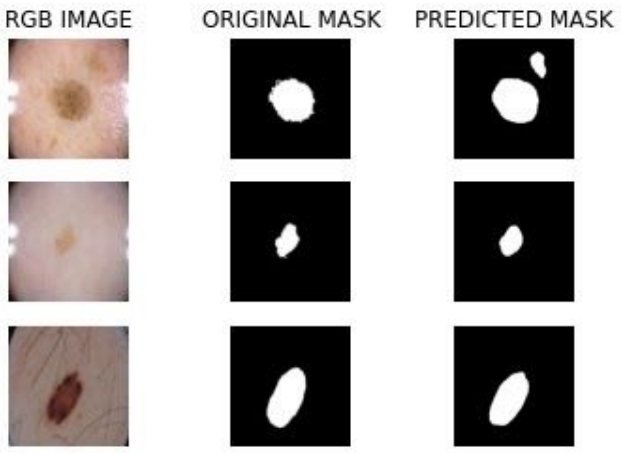


Figure 6.4 Comparison of original and predicted masks from Ternaus Network.

In order to measure the performance of the network, we use the dice coefficient corresponding to the F1 Score metric for segmentation networks. It is the most reliable metric for the evaluation of such networks because after we export the binary masks, the dice coefficient measures the amount of the area in which we will have overlapping between the areas of ground truth and predicted masks.

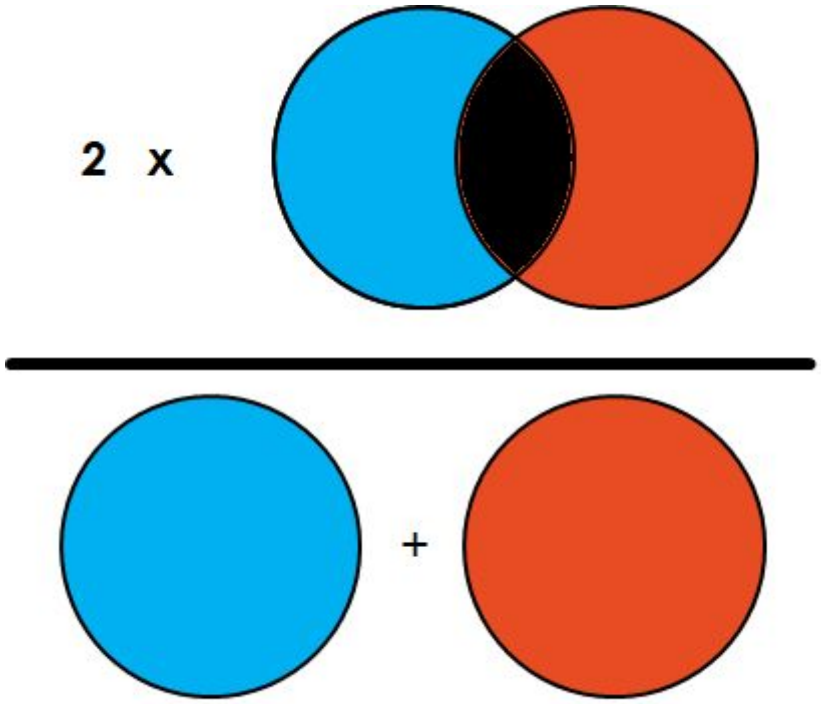


Figure 6.5 Dice Coefficient

In our case, the dice coefficient is approximately ~ 78 % .

6.4 Experiments and Comparison of the Results

As the next and final task, we have to feed our networks with the segmented dataset and then compare the results.

6.4.1 Accuracy Metrics For Segmented Images

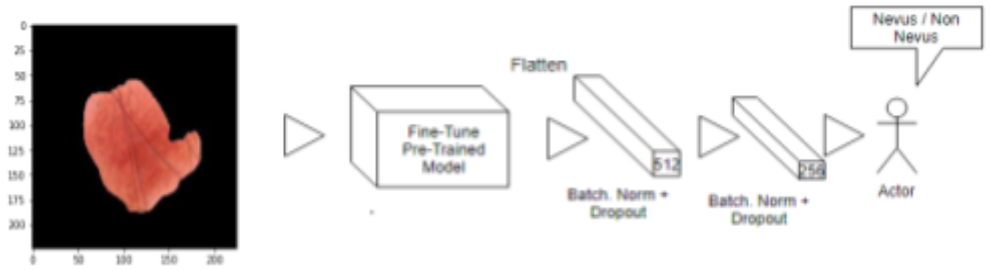


Figure 6.6 Repeat Experiments With Segmented Images

In order to have a framework of validation for both the segmentation and the classification results, we quote the accuracy diagram of each network along with the confusion matrix, and the classification report table, as we did in Chapter 5.

Shallow Network:

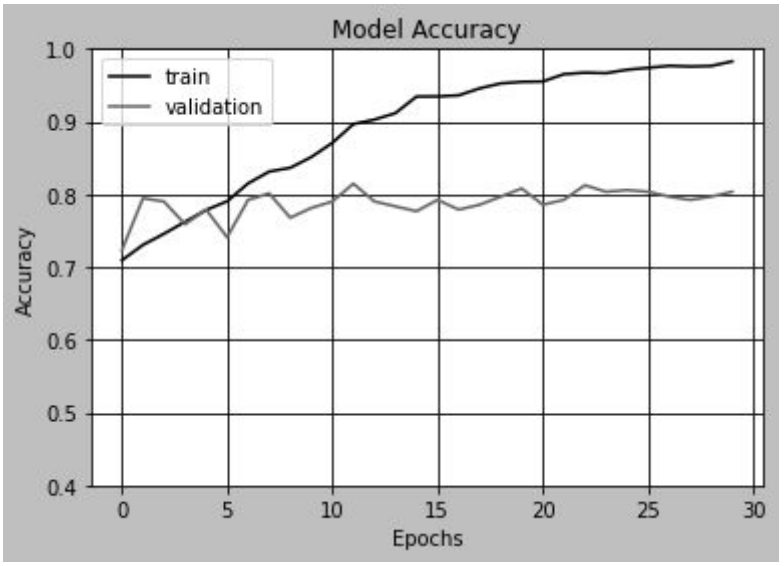


Figure 6.7 Training Process

Segmented Images

Confusion Matrix	Non Nevus	Nevus
Non Nevus	463	137
Nevus	131	469

Table 6.1 Shallow Network Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.78	0.77	0.78
Nevus	0.77	0.78	0.78

Table 6.2 Shallow Network Precision,Recall,F1 Score

In the case of the basic network, we observe that the segmentation did not help to increase its metrics, on the contrary we had a decrease in relation to the corresponding for non-segmented images of the order of 5% on average.

Vgg 19:

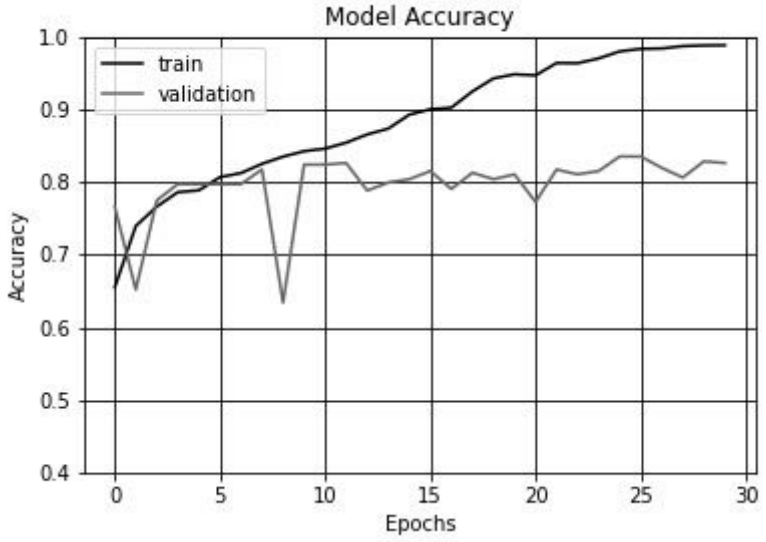


Figure 6.8 Vgg19 Training Process

Segmented Images

Confusion Matrix	Non Nevus	Nevus
Non Nevus	521	79
Nevus	183	417

Table 6.3 Vgg19 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.74	0.87	0.80
Nevus	0.84	0.69	0.76

Table 6.4 Vgg19 Precision,Recall,F1 Score

Regarding Vgg19, its metrics for segmented images showed an even larger drop than in the case of the basic network, of the order of 6 to 7% on average due to the loss of batch normalization from the Vgg19 network, something that will be analyzed below in the final comparison of the networks.

Resnet 50:

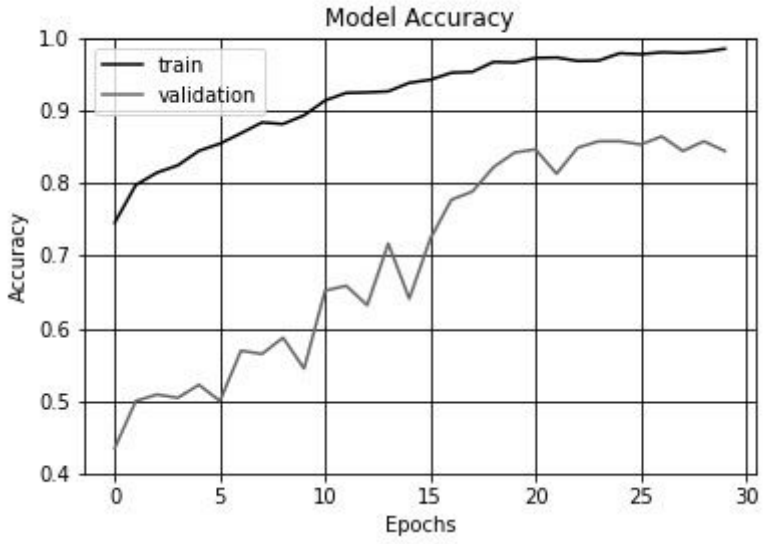


Figure 6.9 Resnet50 Training Process

Segmented Images

Confusion Matrix	Non Nevus	Nevus
Non Nevus	521	79
Nevus	127	473

Table 6.5 Resnet 50 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.80	0.87	0.83
Nevus	0.86	0.79	0.82

Table 6.6 Resnet 50 Precision,Recall,F1 Score

In the case of Resnet 50 we observe for the first time so far an increase of metrics by 3% on average. This is due to the batch normalization which helps in the normalization of the data as models that use it find possible changes that can be made in the datasets because the data is normalized.

Inception V3:

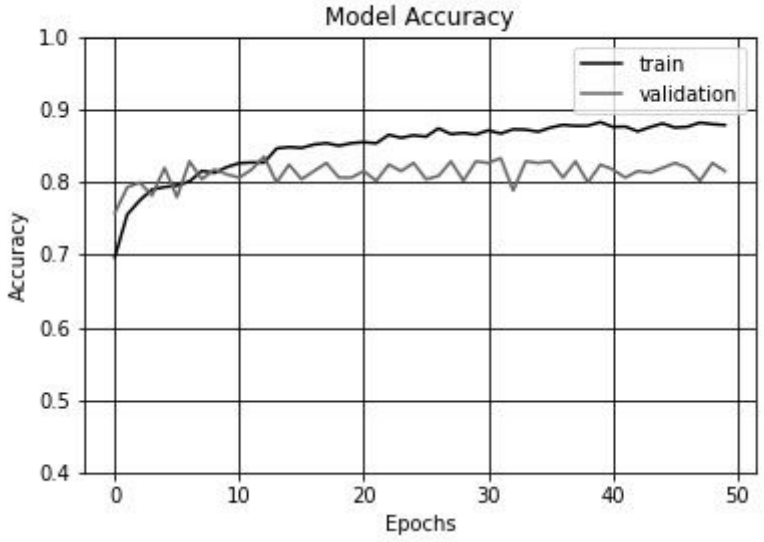


Figure 6.10 Inception V3 Training Process

Segmented Images

Confusion Matrix	Non Nevus	Nevus
Non Nevus	525	75
Nevus	150	450

Table 6.7 Inception V3 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.78	0.88	0.82
Nevus	0.86	0.75	0.80

Table 6.8 Inception V3 Precision,Recall,F1 Score

Regarding the case of Inception V3, which had the best metrics in our first approach to the problem, we observe here a drop of 10% on average which is the largest drop we have seen on a network so far.

This is due as for Vgg19 mainly to the absence of batch normalization layer and will be analyzed as for Vgg19 in the final comparison of the chapter.

Xception:

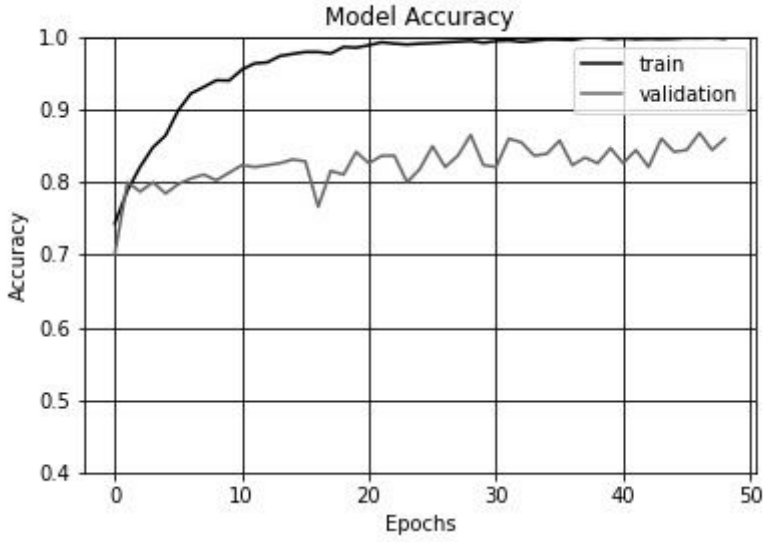


Figure 6.11 Xception Training Process

Segmented Images

Confusion Matrix	Non Nevus	Nevus
Non Nevus	521	79
Nevus	127	473

Table 6.9 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.80	0.87	0.83
Nevus	0.86	0.79	0.82

Table 6.10 Precision,Recall,F1 Score

A small drop in the metrics (3% on average) as well as a better performance in relation to the other networks excluding the Resnet 50, we observe in the Xception model given that here too a batch normalization layer is used as well as depthwise separable convolutions.

MobileNet V1:

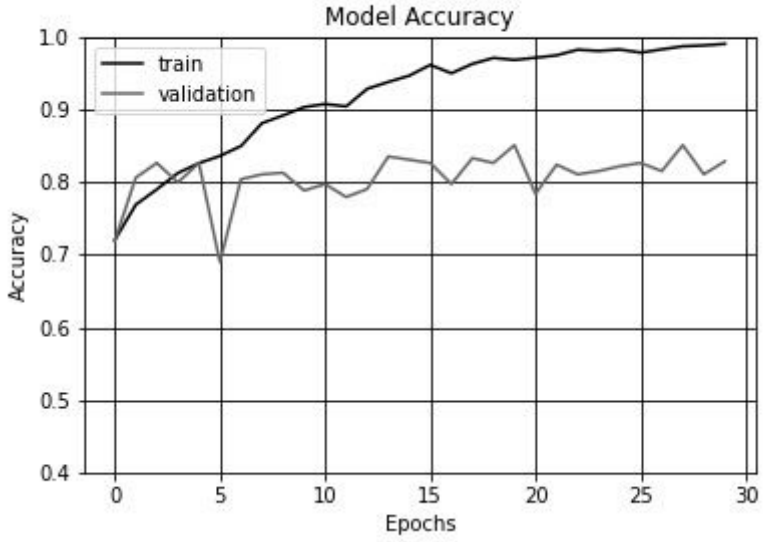


Figure 6.12 MobileNet V1 Training Process

Segmented Images

Confusion Matrix	Non Nevus	Nevus
Non Nevus	418	182
Nevus	55	545

Table 6.11 MobileNet V1 Confusion Matrix

Classification report

	Precision	Recall	F1 Score
Non Nevus	0.88	0.70	0.78
Nevus	0.75	0.91	0.82

Table 6.12 MobileNet V1 Precision,Recall,F1 Score

Finally, as far as the Mobile Net V1 is concerned, we have an average drop of 8% in the metric and best efficiencies between the networks for segmented images with the exception of Resnet 50 and Xception.
This is because on the one hand it has batch normalization layers and also uses depthwise separable convolutions, on the other hand it does not use the shortcuts that we find on the Resnet 50 or the parallel convulsions that the Xception uses to have enriched information.

6.4.2 F1 Score Comparison and Discussion

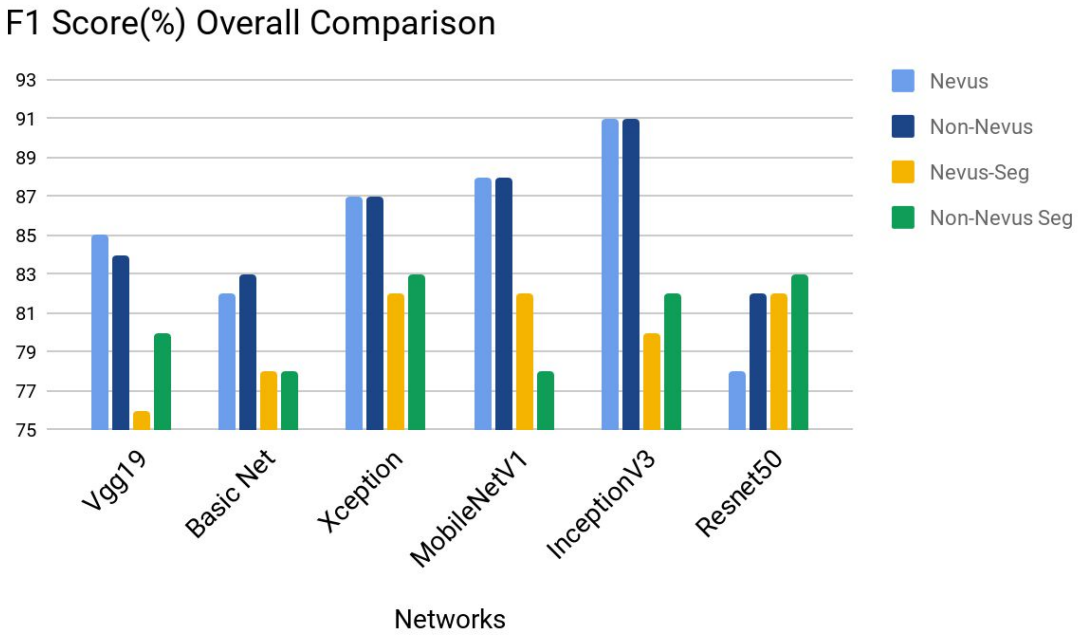


Figure 6.13 F1 Comparison With Segmented and Non Segmented Images

The bar chart shows the F1 Scores of six networks with two different classes of segmented images : Shallow net, Vgg19, Resnet 50, Inception V3, Xception and MobileNet V1.

Shallow net has worse performance with segmented images, with 80% accuracy and 78% F1 Score for both nevus and non nevus.
Vgg19 network for segmented images is not as good as with non segmented (82% accuracy with 80% and 76% F1 Score for non nevus and nevus images respectively), since it does not have batch normalization layers. Batch normalization is very important because thanks to that, models are trained much faster and we reduce the covariance shift, i.e. if something on the data changes, the model is capable of finding this change due to the fact all the data is normalized.
Now while InceptionV3 seems to be the best model for non segmented images, we notice that Xception, Resnet 50 and MobilenetV1 are the best models because they apply on dataset, batch normalization and separative convolutions.
Nevertheless we have to make a note:
If we look at the graph, we can see that overall for segmented images we have worse efficiency than with non segmented images.
The reason behind this is that sometimes the binary mask is not perfect and it may hide crucial regions of the skin lesion, and as we remember dice coefficient is 78% for binary masks which means that there is a 22% wrong in predictions.
Also the black background creates a bias that may be crucial to the worse metrics presented by the models with segmented images.
As a result, the best model for our purpose is the InceptionV3 with 91% F1 Score in nevus-non nevus detection, for non segmented images.
Also, same as in subchapter 6.4.2 if we want to take a better look at the network metrics for segmented images, we can observe the following table.

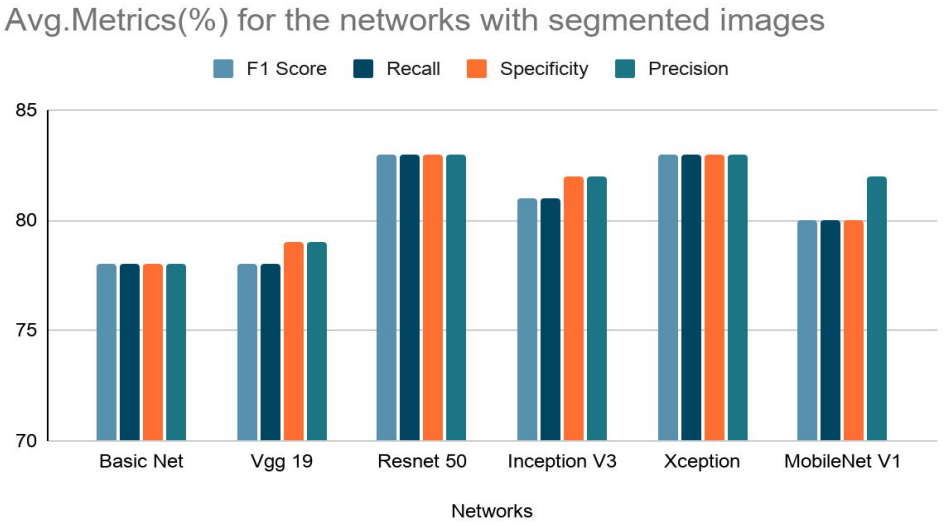


Figure 6.14 Average Network Metrics

As we observe here the metrics are at the same levels, however they are reduced for segmented images, like F1 score for single run.
It is something that should be taken very seriously as metrics like F1 score and specificity are representative of similar diagnostic problems.

6.4.3 Metrics Comparison and Discussion for 10-Fold Cross Validation

In order to have a complete picture of how well our networks operate in real world conditions, we also performed this 10-fold cross validation for the Resnet 50 that had the best metrics for segmented images. Finally we quote ,in addition to the table of metrics,a comparative bar chart and with the Inception V3 for non segmented images from the subchapter 5.5, to see from our two approaches which network responds best to the conditions mentioned above.

Resnet 50	F1 Score	Recall	Specificity
Non Nevus	0.83	0.85	0.85
Nevus	0.81	0.80	0.79

Table 6.13 Cross Validation Avg.Metrics Table

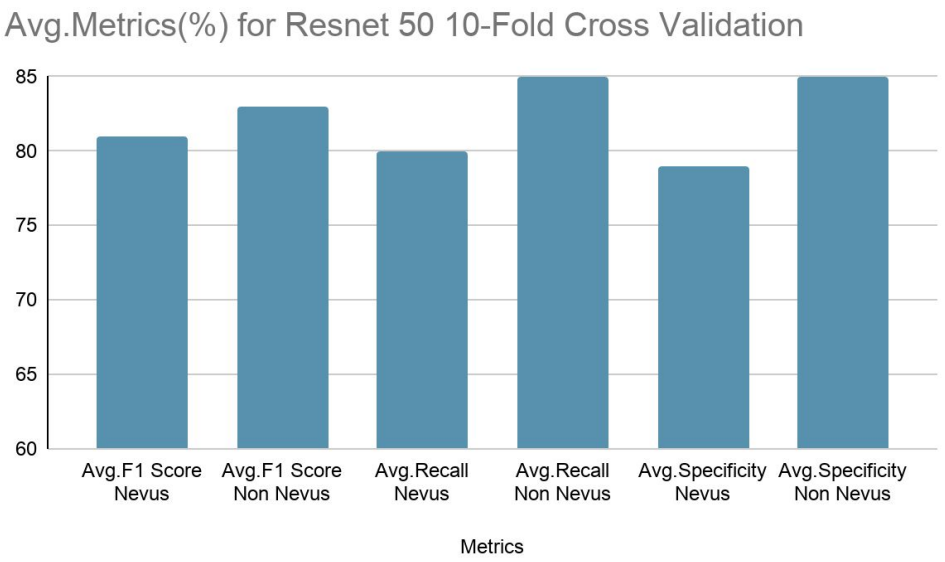


Figure 6.15 Metrics Comparison for Cross Validation Network

As we can see, the measurements here are at the same level for the case of a run, while in terms of the bar chart, we see that here too the Inception V3 approach is the best overall.

Chapter 7
Conclusion and Future Work

Throughout this thesis we studied and compared methods of nevus classification.
In Chapter 1, we made an introduction to the problem and talked about medical image analysis.
In Chapter 2, we talked about dermatoscopes, conventional methods of extracting features, as well as skin detection and hair removal from the region of interest.
In Chapter 3, we made an introduction to basic concepts of machine and deep learning as well as their use in the field of medical image analysis.
We talked about some of the basic building blocks of a network as well as the problems they face and the evaluation metrics we used.
In Chapter 4, we talked about the transfer learning concept as well as the deep learning structures that we used in our case.
In Chapter 5, we presented our first approach to the problem, quoted the data we worked on, and then compared the different network metrics to find the most efficient structure.
In Chapter 6, we presented our second approach where we isolated the skin lesion with segmentation techniques and repeated the evaluation and comparison process between the structures and regarding the second approach and then between the 2 approaches.
In conclusion we have accomplished some important things in this thesis such as, comparison of various pre trained Cnn networks in terms of minimizing the misclassification error. Also comparison of pretrained Cnns combined with Unet feature extraction model, and fine tuning of the above pretrained models. In the end, the end to end approach turned out to be more effective than additional feature extraction techniques, and as a result we managed to classify correctly nevus and non nevus lesions with F1 Score ~91%.
Nevertheless, there is always room for improvement by analyzing and applying different aspects of software such as balancing, hair removal, etc. To decrease diagnostic time of such a system, an interesting point is the developed software to be applicable in real time.
MobileNet is a lightweight engineering system which is progressively reasonable for portable and inserted based vision applications.
Nowadays deep learning is the best solution of skin cancer diseases classification and recognition of cancerous diseases. In the future, MobileNet for skin cancer disease will help medical stakeholders to avoid the conventional lab and in-vivo tests. Without testing or the use of x rays, skin cancer will be automatically detected using the MobileNet algorithm embedded in mobile.

Examples such as (Azzo et al., 2019), (Wibowo et al., 2020) and (Velasco et al., 2019), are projects with prospects on the skin lesion detection sector, for MobileNet applications.

Additionally, remarkable progress is observed in the hyperspectral imaging sector, as they testify projects like (WIREs Authors, 2020), (Johansen et al., 2019) and (Chen et al., 2020).

Taking inspiration from such works and being motivated by optics that we did not consider in this work, as we worked with RGB images an idea is $\tau\ o$ explore how much we can improve the quality of the representation, based on different color spaces of the image, 3D convolutional kernels as we worked with 2D convolutional kernels. Also creating a bounding box that can minimize the region of non interest is a future goal associated with this thesis.

Bibliography

1.Ai in healthcare heatmap: From diagnostics to drug discovery startups, the category heats up. (9/2016).https://www.cbinsights.com/research/artificial-intelligence-healthcare-investment-heatmap/

2.Akram, T., Lodhi, H. M.J., Naqvi, S. R., Naeem, S., Alhaisoni, M., Ali, M., Haider, S. A., & Qadri, N. N. (31/3/2020). *A multilevel features selection framework for skin lesion classification*.Vol.12

3.Ali, A.-R., LI, J., Yang, G., & O'Shea, S. J. (29/6/2020). *A machine learning approach to automatic detection of irregularity in skin lesion border using dermoscopic images*.

Applications - InceptionV3. https://keras.io/applications/ #inceptionv3

4.Argenziano, G., Fabbrocini, G., Carli, P., Giorgi, V. D., Sammarco, E., & Delfino, M. (1998). *Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis.* Archives of Dermatology.Pages 1563-70

5.Argenziano, G., & Soyer, H. P. (7/2001). *Dermoscopy of pigmented skin lesions–a valuable tool for early diagnosis of melanoma*. The Lancet Oncology.Pages 443-9

6.Argenziano, G., Zalaudek, I., Ferrara, G., Johr, R., Langford, D., Puig, S., Soyer, H. P., & Malvehy, J. (1/2007). *Dermoscopy features of melanoma incognito: indications for biopsy*. Journal of the American Academy of Dermatology.Vol 56.Pages 508-513

7.Australian Medical Association et al. (8/2005). *Review of health workforce*. submission to the productivity commission. Canberra, Australia.Pages 8-18,51

8.Azzo, F. A., Awad, A., & Milanova, M. (7/2019). Skin Lesion Detection by Android Camera based on SSD-Mo- bilenet and TensorFlow Object Detection API (A. M. Taqi, Ed.).

Pages 6-12

9.Bafounta, M. L., Beauchet, A., Aegerter, P., & Saiag, P. (10/2001). *Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma?: Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests*. Archives of dermatology.Pages 1343-50

10.Barata, C., Marques, J. S., & Rozeira, J. (4/2011). Detecting the pigment network in dermoscopy images: A directional approach. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (pages 5120–5123). IEEE.

11.Barata, C., Ruela, M., Francisco, M., Mendonca, T., & Marques, J. S. (9/2014). *Two systems for the detection of melanomas in dermoscopy images using texture and color features*. IEEE Systems Journal.Vol.8.Pages 965-979

12.Binder, M., Schwarz, M. P., Steiner, A., Kittler, H., Muellner, M., Wolff, K., & Pehamberger, H. (2/1997). *Epiluminescence microscopy of small pigmented skin lesions: short-term formal training improves the diagnostic performance of dermatologists*. Journal of the American Academy of Dermatology.Vol.36.Pages 197-202

13.Binder, M., Steiner, A., Schwarz, M., Knollmayer, S., Wolff, K., & Pehamberger, H. (4/1994). *Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study*. British Journal of Dermatology.Vol.130.Pages 460-465

14.Blum, A., Luedtke, H., Ellwanger, U., Schwabe, R., Rassner, G., & Garbe, C. (11/2004). *Digital image analysis for diagnosis of cutaneous melanoma. development of a highly effective computer algorithm based on analysis of 837 melanocytic lesions*. British Journal of Dermatology.Pages 1029-38

15.Brown, K. (10/2015). *Shortage of dermatologists in public hospitals*.

16.Burges, C. J. (6/1998). A tutorial on support vector machines for pattern recognition. In *Data mining and knowledge discovery*.Vol.2.Pages 121-167

17.Caputo, B., Panichelli, V., & Gigante, G. (2/2002). Toward a quantitative analysis of skin lesion images. In *Studies in Health Technology and Informatics* (pages 509–513).

18.Celebi, M. E., Iyatomi, H., Stoecker, W. V., Moss, R. H., Rabinovitz, H. S., Argenziano, G., & Soyer, H. P. (9/2008). *Automatic detection of blue-white veil and related structures in dermoscopy images*. Computerized Medical Imaging and Graphics.Pages 670-77

19.Chen, J., Wang, X., Wu, Q., & Mo, J. (10/2020). *Skin melanoma detection based on hyperspectral imaging and deep-learning techniques*.Vol.11553

20.Chockley, K., & Emanuel, E. (12/2016). The end of radiology? Three threats to the future practice of radiology. *Journal of the American College of Radiology*.Pages 1415-20

21.Chollet, F. (2017). *Deep Learning with Python* (1 Edition ed.). Manning, Shelter Island, NY.

22.Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., & Smith, J. R. (10/2015). Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging* .Vol.9352.(pages 118–126). Springer.

23.Codella, N., Nguyen, Q. B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., & Smith, J. R. (10/2016). *Deep learning ensembles for melanoma recognition in dermoscopy images*. arXiv preprint arXiv:1610.04662.Vol.2

24.Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., & Kittler, H. (10/2017). *Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)*. arXiv preprint arXiv:1710.05006.Vol.3

25.D,F.(10/2017).TowardsDataScience.

https://towardsdatascience.com/batch-normalization-in-neural-networks-1ac91516821c

26.Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei, L. F. (6/2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pages 248–255). IEEE.

27.Dermnet - Skin Disease Atlas. http://www.dermnet.com/

28.Di Leo, G., Liguori, C., Paolillo, A., & Sommella, P. (2008). An improved procedure for the automatic detection of dermoscopic structures in digital elm images of skin lesions. In *Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008. VECIMS* (pages 190–194). IEEE.

29.Di Leo, G., Paolillo, A., Sommella, P., & Fabbrocini, G. (1/2010). Automatic diagnosis of melanoma: a software system based on the 7-point check-list. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference* (pages 1–10). IEEE.

30.Dreiseitl, S., Machado, L. O., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2/2001). *A comparison of machine learning methods for the diagnosis of pigmented skin lesions*. Journal of Biomedical Informatics.Pages 28-36

31.Edinburgh Research and Innovation. Dermofit Image Library. https: //licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html

32.Ercal, F., Chawla, A., Stoecker, W. V., Lee, H. C., & Moss, R. H. (9/1994). Neural network diagnosis of malignant melanoma from color images. In *IEEE Transactions on biomedical engineering*.Vol.41.Pages 837-845

33.Esteva, A., Kurpel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (1/2017). *Dermatologist-level classification of skin cancer with deep neural networks*. Nature.Pages 115-118

34.Fabbrocini, G., Betta, G., DiLeo, G., Liguori, C., Paolillo, A., Pietrosanto, A., Sommella, P., Rescigno, O., Cacciapuoti, S., & Pastore, F. (11/2010). *Epiluminescence image processing for melanocytic skin lesion diagnosis based on 7-point checklist: a preliminary discussion on three parameters*. The open dermatology journal.Vol.4.Pages 110-115

35.Faziloglu, Y., Stanley, R. J., Moss, R. H., Van Stoecker, W., & McLean, R. P. (6/2003). *Colour histogram analysis for melanoma discrimination in clinical images*. Skin Research and Technology.Vol.9.Pages 147-56

36.Fleming, M. G., Steger, C., Zhang, J., Gao, J., Cognetta, A. B., & Dyer, C. R.(Sept-Oct/1998). *Techniques for a structural analysis of dermatoscopic imagery*. Computerized medical imaging and graphics.Pages 375-89

37.Freedberg, K. A., Geller, A. C., Miller, D. R., Lew, R. A., & Koh, H. K. (11/1999). *Screening for malignant melanoma: A cost-effectiveness analysis*. Journal of the American Academy of Dermatology.Pages 738-45

38.Freund, Y., & Schapire, R. E. (8/1997). *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of computer and system sciences.Vol.55.Pages 119-39

39.Gachon, J., Beaulieu, P., Sei, J. F., Gouvernet, J., Claudel, J. P., Lemaitre, M., Richard, M. A., & Grob, J. J. (4/2005). *First prospective study of the recognition process of melanoma in dermatological practice*. Archives of Dermatology.434-8

40.Garnavi, R., Aldeen, M., & Bailey, J. (11/2012). *Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis*. IEEE Transactions on Information Technology in Biomedicine.Pages 1239-52

41.Goodson, A. G., & Grossman, D. (5/2009). *Strategies for early melanoma detection: Approaches to the patient with nevi*. Journal of the American Academy of Dermatology.Pages 719-35

42.Zendesk.https://www.zendesk.com/blog/machine-learning-and-deep-learning/

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy,, A., Venugopalan,, S., Widner, K., Madams, T., & Cuadros, J. (12/2016). *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs*. Jama.Pages 2402-2410

43.Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., & Halpern, A. (5/2016). *Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI)* (arXiv preprint arXiv:1605.01397 ed.). International Skin Imaging Collaboration (ISIC).

44.Guy, G. P., Machlin, S. R., Ekwueme, D. U., & Yabroff, K. R. (2/2015). *Prevalence and costs of skin cancer treatment in the US, 2002- 2006 and 2007- 2011*. American Journal of Preventive Medicine.Pages 183-187

45.Guy Jr, G. P., & Ekwueme, D. U. (10/2011). *Years of potential life lost and indirect costs of melanoma and non-melanoma skin cancer*. Pharmacoeconomics,Pages 863-74

46.He, K., Zhang, X., Ren, S., & Sun, J. (12/2015). Deep Residual Learning for Image Recognition. In *Technical report*.

47.He, K., Zhang, X., Ren, S., & Sun, J. (12/2015). *Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385*.

48.Henning, J. S., Dusza, S. W., Wang, S. Q., Marghoob, A. A., Rabinovitz, H. S., Polsky, D., & Kopf, A. W. (1/2007). *The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy*. Journal of the American Academy of Dermatology.

49.Holzinger, A., Kieseberg, P., Tjoa, A. M., & We, E. R. (2019). *Machine Learning and Knowledge Extraction*.Pages 45-52

50.Holzinger, A., Kieseberg, P., Tjoa, A. M., Weippl, E., & Liang, X. (8/2019). In *Machine Learning and Knowledge Extraction: Third IFIP* (pages 381-382).

51.Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2019, May 21). Classification of skin lesions using transfer learning and augmentation with Alex-net.

52.IEEE International Symposium on Biomedical Imaging. (2016). http:// biomedicalimaging.org/

53.Iglovikov, V., & Shvets, A. (1/2018). *TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation*.

54.ISIC Archive - International Skin Imaging Collaboration: Melanoma Project. (2016). https://isic-archive.com/

55.ISIC. ISBI 2016-17: Skin lesion analysis towards melanoma detection. https://goo.gl/t96E77

56.ISIC. ISIC skin image analysis workshop and challenge @ miccai 2018. (2017). https://goo.gl/aRxnBt

57.Jha, S. (2016). *Will Computers Replace Radiologists?* http://www.medscape.com/viewarticle/863127,

58.Jia, Y., Shellhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (11/2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia* (pages 675–678). ACM.Pages 675-678

59.Johansen, T. H., Mollersen, K., Ortega, S., Fabelo, H., Garcia, A., Callico, G. M., & Godtliebsen, F. (4/2019). Recent advances in hyperspectral imaging for melanoma detection.

60.Kaushik, A. (2020). *Understanding the VGG19 Architecture*. OpenGenus IQ. https://iq.opengenus.org/vgg19-architecture/

61.Kawahara, J., BenTaieb, A., & Hamarneh, G. (2016). Deep features to classify skin lesions. In *IEEE International Symposium on Biomedical Imaging (IEEE ISBI), Prague, Czech Republic* (pages 1397–1400). IEEE.

62.Kawahara, J., & Hamarneh, G. (2016). Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In *In International Workshop on Machine Learning in Medical Imaging* (pages 164–171). Springer.

63.Khakabi, S., Wighton, P., Lee, T. K., & Atkins, M. S. (2/2012). *Multi-level feature extraction for skin lesion segmentation in dermoscopic images*.

64.Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. (3/2002). *Diagnostic accuracy of dermoscopy*. The Lancet Oncology.Pages 159-65

65.Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (volume 14 ed., pages 1137–1145).

66.Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (pages 1097–1105). Curran Associates, Inc., Red Hook, New York, USA.

67.Lakhami, P., & Sundaram, B. (4/2017). *Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks*. Radiology.Vol.284

68.LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (12/1998). *Gradient-based learning applied to document recognition*. Proceedings of the IEEE.Vol.86.Pages 2278-2324

69.Liao, H. (2016). *A deep learning approach to universal skin disease classification. University of Rochester Department of Computer Science*. CSC.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sanchez, C. I. (2/2017). *A survey on deep learning in medical image analysis*. arXiv preprint arXiv:1702.05747.Vol.2.Pages 60-88

70.Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., & Corrado, G. S. (3/2017). *Detecting cancer metastases on gigapixel pathology images*. arXiv preprint arXiv:1703.02442.Vol.2

71.Lopez, A. R. (1/2017). *SKIN LESION DETECTION FROM DERMOSCOPIC IMAGES USING CONVOLUTIONAL NEURAL NETWORKS*. Universitat Politècnica de Catalunya.

Lopez, A. R., Giro-i-Nieto, X., Burdick, J., & Marques, O. (2/2017). *Skin Lesion Classification from Dermoscopic Images Using Deep Learning Techniques*.

72.Mallory, P. (2016, October). *Your smartphone as a dermatologist: Fast, cheap...and often wrong*. https://www.kqed.org/futureofyou/18178/

your-smartphone-as-dermatologist-fast-cheap-and-often-wrong

73.Marchetti, M. A., Codella, N. C., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., Mishra, N., Carrera, C., Celebi, M. E., & DeFazio, J. L. (2/2018). *Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy*

of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. Journal of the American Academy of Dermatology.Pages 270-277

74.Menard, S. (2018). *Applied logistic regression analysis, volume 106*. SAGE publications.

Mendonca, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., & Rozeira, J. (2013). Ph 2-a dermoscopic image database for research and benchmarking. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (pages 5437–5440). IEEE.

75.Menzies, S. W., Crotty, K. A., Ingvar, C., & McCarthy, W. H. (10/2002). *An atlas of surface microscopy of pigmented skin lesions: dermoscopy*. McGraw-Hill Sydney, Australia.

Menzies, S. W., & Zalaudek, I. (9/2006). *Why perform dermoscopy?: The evidence for its role in the routine management of pigmented skin lesions*. Archives of dermatology.Pages 1211-1212

76.Mukherjee., S. (2017, April). A.I. versus M.D.: What happens when diagnosis is automated? *The New Yorker*.

77.Murzaku, E. C., Hayan, S., & Rao, B. K. (8/2014). *Methods and rates of dermoscopy usage: a cross-sectional survey of us dermatologists stratified by years in practice*. Journal of the American Academy of Dermatology.vol.71.Pages 393-395

78.Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., Bilek, P., Falco, O. B., & Plewig, G. (4/1994). *The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions*. Journal of the American Academy of Dermatology.Pages 551-559

79.Obermeyer, Z., & Emanuel, E. J. (9/2016). *Predicting the future big data, machine learning, and clinical medicine*. The New England journal of medicine.Pages 1216-1219

80.Obermeyer, Z., & Lee, T. H. (9/2017). *Lost in thought the limits of the human mind and the future of medicine*. New England Journal of Medicine.1209-1211

81.Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (3/2018). *The building blocks of interpretability*. Distill. https://distill.pub/2018/building-blocks

82.Ozkan, I. A., & Koklu, M. (12/2017). *Skin Lesion Classification using Machine Learning Algorithms*.Pages 285-289

83.Page, E. H. (2/2019). *Description of skin lesions*. https://goo.gl/m9ybFp

84.Pan, S. J., & Yang, Q. (10/2010). A survey on transfer learning. In *IEEE Transactions on knowledge and data engineering*. IEEE.Vol.22.Pages 1345-1359

85.Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2/2018). *Multimodal explanations: Justifying decisions and pointing to the evidence*. arXiv preprint arXiv:1802.08129.

86.Pathan, S., Prabhu, K. G., & Siddalingaswamy., P. (1/2018). *Techniques and algorithms for computer aided diagnosis of pigmented skin lesions a review*. Biomedical Signal Processing and Control.Vol.39.Pages 237-262

87.Pehamberger, H., Steiner, A., & Wolff, K. (10/1987). *In vivo epiluminescence microscopy of pigmented skin lesions. i. pattern analysis of pigmented skin lesions*. Journal of the American Academy of Dermatology.Vol.17.Pages 571-583

88.Pellacani, G., Grana, C., Cucchiara, R., & Seidenari, S. (7/2004). *Automated extraction and description of dark areas in surface microscopy melanocytic lesion images*. Dermatology.Vol.208.Pages 21-26

89.Piccolo, D., Ferrari, A., Peris, K., Daidone, R., Ruggeri, B., & Chimenti, S. (9/2002). *Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: a comparative study*. British Journal of Dermatology.Vol.147.Pages 481-486

90.Poplin, R., Varadarajan, A. v., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., Webster, J. Y., & Dale, R. (3/2018). *Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning*.Vol.2.Pages 158-164

91.Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., & Shpanskaya,, K. (12/2017). *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning*. arXiv preprint arXiv:1711.05225.Vol.3

92.Rastgoo, M., Morel, O., Marzani, F., & Garcia, R. (2015). Ensemble approach for differentiation of malignant melanoma. In *Twelfth International Conference on Quality Control by Artificial Vision 2015, volume 9534,* (page 953415). International Society for Optics and Photonics.

93.Ratul, A. R., Mozaffari, M. H., Parimbelli, E., & Lee, W. (8/2019). *Atrous Convolution with Transfer Learning for Skin Lesions Classification*.

94.Ribeiro, M. T., Singh, S., & Guestrin, C. (8/2016). *Why should i trust you?: Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.Pages 1135-1144

95.Rigel, D. S., Friedman, R. J., Kopf, A. W., & Polsky, D. (8/2005). *Abcde an evolving concept in the early detection of melanoma*. Archives of dermatology.Vol.141.Pages 1032-1034

96.Ronneberger, O., Fischer, P., & Brox, T. (5/2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.

97.Rumelhart, D. E., McClelland, J. L., & Group, P. R. (7/1987). *Parallel distributed processing, volume 1*. MIT press Cambridge, MA.Vol.1.

98.Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (4/2015). *ImageNet large scale visual recognition challenge*. International Journal of Computer Vision.Vol.115.Pages 211-252

99.Sadeghi, M., Razmara, M., Lee, T. K., & Atkins, M. S. (3/2011). *A novel method for detection of pigment network in dermoscopic images using graphs.Computerized Medical Imaging and Graphics*.Vol.35.Pages 137-143

100.Safavian, S. R., & Landgrebe, D. (May-June/1991). A survey of decision tree classifier methodology. In *IEEE transactions on systems, man, and cybernetics*. IEEE.Vol.21.Pages 660-674

101.Sagar, A., & J, D. (5/2020). *Convolutional Neural Networks for Classifying Melanoma Images*.

102.Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (6/2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Technical report*.

103.Schaefer, G., Krawczyk, B., Celebi, M. E., & Iyatomi, H. (12/2014). An ensemble classification approach for melanoma diagnosis. In *Memetic Computing*.Vol.6.Pages 233-240

104.Seidenari, S., Pellacani, G., & Grana, C. (12/2005). Pigment distribution in melanocytic lesion images: a digital parameter to be employed for computer-aided diagnosis. In *Skin Research and Technology*.Vol.4.Pages 236-241

105.Sheha, M. A., Mabrouk, M. S., & Sharawy, A. (3/2012). Automatic detection of melanoma skin cancer using texture analysis. In *International Journal of Computer Applications*.Vol.42.Pages 22-26

106.Siegel, R. L., Miller, K. D., & Jemal, A. (Jan-Feb/2018). *Cancer statistics*. CA: a cancer journal for clinicians.Vol.68.Pages 7-30

107.Simonyan, K., & Zisserman, A. (9/2014). *Very deep convolutional networks for large scale image recognition*. arXiv preprint arXiv:1409.1556.Vol.6

108.Spiegel, R. L., Miller, K. D., & Jemal, A. (Jan-Feb/2016). *Cancer statistics*. CA: A Cancer Journal for Clinicians.Vol.66.Pages 7-30

109.Stanley, R. J., Stoecker, W. V., & Moss, R. H. (2/2007). *A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images*. Skin Research and Technology.Vol.13.Pages 62-72

110.Stoecker, W. V., Gupta, K., Stanley, R. J., Moss, R. H., & Shrestha, B. (8/2005). Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color. In *Skin Research and Technology*.Vol.11.Pages 179-184

111.Stoecker, W. V., Wronkiewiecz, M., Chowdhury, R., Stanley, R. J., Xu, J., Bangert, A., Shrestha, B., Calcara, D. A., Rabinovitz, H. S., & Oliviero, M. (3/2011). Detection of granularity in dermoscopy

images of malignant melanoma using color and texture features. In *Computerized Medical Imaging and Graphics*.Vol.35.Pages 144-147

112.Stolz, W., Holzel, D., Riemann, A., Abmayr, W., Przetak, C., Bilek, P., Landthaler, M., & Falco, O. B. (1991). *Multivariate analysis of criteria given by dermatoscopy for the recognition of melanocytic lesions*. Book of Abstracts, Fiftieth Meeting of the American Academy of Dermatology, Dallas.Vol.42.Pages 77-83

112.Stolz, W., & Kunz, M. *ABCD rule*. Dermoscopedia. https://dermoscopedia.org/ABCD_rule

113.Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2/2016). Inceptionv4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Technical report*.

114.Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (12/2015). *Rethinking the Inception Architecture for Computer Vision*.Vol.3

115.Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking theInception architecture for computer vision. In *arXiv preprint arXiv:1512.00567* (pages 2818–2826).

116.Tabatabaie, K., Esteki, A., & Toossi, P. (11/2009). *Extraction of skin lesion texture features based on independent component analysis*.Vol.15.Pages 433-439

117.Tascon, S. (2019). *Dermoscopy images with two classes: nevus and malignant*. Kaggle. https://www.kaggle.com/sergio814/dermoscopy-images

118.Velasco, J., Pascion, C., Alberio, J. W., Apuang, J., Cruz, J. S., Gomez, M. A., Molina, B. J., Tuala, L., Thio-ac, A., & Jorda, R. J. (10/2019). A Smartphone-Based Skin Disease Classification Using MobileNet CNN.Vol.8.Pages 2632-2637

119.Vestergaard, M., Macaskill, P., Holt, P., & Menzies, S. (9/2008). *Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: A metaanalysis of studies performed in a clinical setting*. British Journal of Dermatology.Vol.159.Pages 669-676

120.Walter, M. (9/2016). *Is this the end? Machine learning and 2 other threats to radiology's future*. goo.gl/M9X3SF

121.Weese, J., & Lorenz, C. (10/2016). *Four challenges in medical image analysis from an industrial perspective*. Medical image analysis.Vol.33.Pages 44-49

122.Wibowo, A., Hartanto, C. A., & Wirawan, P. W. (2020, July). Android skin cancer detection and classification based on MobileNet v2 model.Vol.6.Pages.135-148.

123.WIREs Authors. (4/2020). *Detecting skin cancer using hyperspectral images*. Advance Science News.

https://www.advancedsciencenews.com/detecting-skin-cancer-using-hyperspectral-images/.Vol.12

124.Yoshino, S., Tanaka, T., Tanaka, M., & Oka, H. (2004). Application of morphology for detection of dots in tumors. In *SICE 2004 Annual Conference, volume 1,* (pages 591–594). IEEE.

125.Yuan, X., Yang, Z., Zouridakis, G., & Mullami, N. (2006). Svm-based texture classification and application to early melanoma detection. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE* (pages 4775–4778). IEEE.