University of Bergen
Master Thesis in Information Science

Melanoma Classification in Low Resolution Dermoscopy Images Using Deep Learning

By:

Christian Andersen

Supervisor:

Csaba Veres

1 June 2019
Academic Year 2018-2019

Table of contents

1. Abstract

Malignant melanoma is a type of cancer that most commonly affects the surface of the skin, and even though it only makes up a meager 4% of all existing skin cancers, melanoma is culpable for 80% of the deaths caused by all cutaneous cancers[12]. If melanoma is detected in its early stages, it is almost always treated successfully, but this treatment becomes significantly more complicated if it is allowed to develop[2]. Due to overlapping and heavy variations of present artifacts in benign and malignant lesions, real world professionals cannot always provide a reliable diagnosis. As a result of this, leveraging computer aided systems in order to aid in diagnosis of skin cancer has become a very appealing area in recent times. This thesis was based on using deep learning techniques to distinguish between malignant melanoma and benign lesions. A comparative study of the effect three pre-processing steps; hair removal, contrast enhancement and a median filter exert on diagnosing malignant melanoma in low resolution dermoscopy images, was conducted. This effect was then compared to a baseline experiment consisting of dermoscopy images without the use of any manipulation, besides resizing. Cutting edge general object recognition CNNs and a proposed 6-layered CNN, are used and evaluated for the binary task of melanoma recognition. Observations from the findings, exhibited employing contrast enhancement achieved the best result from the proposed model with an accuracy, sensitivity and specificity of 81.25%, 74.7% and 87.8% respectively. The leading object recognition model achieved a sensitivity of 96.64% and a specificity of 46.09%. The proposed CNN was then compared to existing state-of-the-art research incorporating convolutional neural networks for classification. The final study demonstrates that near-to state-of-the-art results are achievable using the proposed network with low resolution dermoscopy images.

2. Introduction

Malignant melanoma is reported to be the most common type of skin cancer as well as being considered one of the most prominent causes of death among young people. This disease was once considered an uncommon occurrence, but has since, over the past decades, increased drastically. This affliction, which tends to originate in the outer layers of the skin before ulceration(penetration of the skin) takes effect, continues to increase its presence globally[2].
Out of all the skin cancers, melanoma is the one that contributes to most deaths annually. This cutaneous cancer develops from melanocytic(pigmented cells) and is almost always curable if detected in its early stages, but becomes much more difficult to treat and dangerous if allowed to spread beyond the point of origin into the body, and as a result the survival rate drops dramatically. Therefore, detecting malignant melanoma as early as possible is vital in order for it to be successfully treated[2].

The arrival of dermoscopy, also known as epiluminoscopy, has contributed to a powerful increase in clinical diagnostic abilities, allowing for a better detection rate of melanoma[2]. Dermoscopy is a method for collecting magnified, illuminated and very clear images of lesions affecting the surface of the skin. Large datasets of dermoscopy images has accumulated as a result of accepting this technology globally due to being widely used by professionals. Paired with new advances in computational power and machine- and deep learning for image recognition, the combination of these two technologies have shown great promise in preliminary trials[2]. The substantial ongoing increases in computational power and ability to access better and larger data sets have had a tremendously positive effect on deep learning neural networks' performances as they have, in the recent years, surpassed human capabilities in image and object classification tasks. One good example of this occurrence being the winners of the widely renowned ImageNet competition between 2015-2017, with the winning network in 2017 achieving a classification rate of 97.3% on the ImageNet data set [16][17].

Another instance of the growth in visual classification tasks worth mentioning, is the case of the winning submission for the german traffic sign recognition benchmark, achieving an accuracy of 99.46% in recognizing german traffic signs, outperforming the recognition rate of that of a human(98.84%)[18].

Creating a computer aided diagnosis tool for predicting skin cancer to assist professionals has been a very appealing area of interest for researchers for quite some time and still is, especially when considering the incredible impact computer vision has had on predicting objects in images recently and what can be accomplished with it. In this field, there are multiple different methods and approaches that can be used to diagnose malignant melanoma, and the ones considered in this thesis will predominantly encompass different convolutional neural networks(CNNs). In recent years, CNNs have been gaining popularity as they have consistently been performing at the very top of image recognition tasks and proven themselves to be very effective at this. CNNs have since, seen a steady increase of research in its field for these types of tasks[14], especially after the breakthrough algorithm winner AlexNet[32]. AlexNet was presented in 2012 at the ImageNet competition which completely outperformed the next closest submission with an error rate improvement of more than 10.8 percentage points[33]. AlexNet remained at the top until it was beaten by another CNN(ResNet-50[34]) produced by Microsoft in 2015[33].

This master thesis is based on the binary classification problem of diagnosing malignant melanoma from benign lesions using supervised learning. Multiple comparative studies are conducted through side by side comparison of the effects different image pre-processing steps have on the classification performance of melanoma. Evaluations of a custom and state-of-the-art CNNs, where the latter have proven to be exceptional at general object recognition tasks in images, are conducted for this task. A final evaluation of the best result I achieve is then compared to other existing related research that deals with the detection of skin cancer using CNNs. I am conducting this research in order to gain knowledge and insight, as well as provide an overview in terms of the aforementioned factors, to find out what contributes to improve or decrease the

performance of models for this specific task. All models that are used to conduct these experiments with, will be trained and tested using low resolution dermoscopy images. The reason behind employing low resolution imagery is mainly attributed to the constraints I have when using a personal computer. These CNNs become much more computationally expensive to train as the dimensions of the images increase. Alternatives that have been considered have been to use proprietary computers at the University of Bergen, but some of these networks I use will be considerably complex and deep, requiring multiple full days to train. The second option was to make use of virtual machines that possessed the necessary specifications to be able to handle these models efficiently, however this may have been an economically costly option. A benefit from employing lower resolution images is the increase in speed when training the networks. Curiosity poses the secondary reason for employing low resolution images in these experiments, with the goal of observing if it is possible to achieve near-to or similar results to existing CNNs that deal with these types of problems and uses images of higher resolutions.

This paper is organized into different sections as follows; After this brief overview of the purpose of this research and background history of the field, the research questions will be presented, followed by different works of research that has been conducted, which directly relates to some part of the research in this thesis. I then explain how the data I used for training and testing the deep learning networks was acquired and the composition of the data set itself. Following this, I delve into the methods of the research where the approach and methods used are presented, what networks are used and how they are evaluated. Techniques that are employed on the data and models to reduce overfitting are described as well as further elaboration on the structure of the deep learning networks used for prediction. Results of performances of the networks used are then presented, evaluated and discussed along with any other noteworthy points, before finally presenting a conclusion to this thesis.

3. Research Questions

Using computer vision for classifying skin cancer is in itself a very challenging task, but another task that is also a difficult one that is closely related is the autonomous segmentation of lesions. Ideally, when providing training images to a network for this task when the goal is to be as accurate as possible, we would like to omit as much as possible that is not of interest such as the normal skin that surrounds the lesion by mapping out and drawing a border as precisely as possible around the area of interest. However, a variety of different artifacts that may hinder this step may be present in an image and pre-processing steps can either on their own or in combination be used to negate some of these effects to a certain degree, in order to make the drawn border more accurate. Plenty of research discuss different techniques and algorithms that can be used for this problem with varying accuracies and employing different pre-processing methods for the data it is designed for, but this thesis will not attempt to tackle the segmentation problem and instead take into consideration some of these image pre-processing steps which leads to the first question I would like to give an answer to in this paper. Three pre-processing steps are used in my research; hair removal, contrast enhancement and a median filter. In the case of hair removal, there has been conducted considerable research in the field by employing this technique as a preliminary step in combination or alone with others for facilitating the separation algorithm of a lesion from the normal surrounding skin[2, 53, 76-78]. A comparative study has also been completed on how well different automatic hair detection and removal algorithms performed[54], but this is solely focused on the performance of how effectively and accurately these algorithms detect and remove occluding hairs in dermoscopy images. To my knowledge, there has not been completed any research that specifically takes into account the direct correlation of how the removal of occluding hairs affect the performance of melanoma recognition as opposed to melanoma recognition without this step. One of the remaining two steps, a median filter, fall under the same category as hair removal in the sense of providing a comparative study of how

these pre-processing steps, both alone or in combination, affect the performance of models compared to a baseline consisting of raw, untouched dermoscopy images. C. Barata et al[22] conducted a comparative study of contrast enhancement techniques, but they used an average high resolution of 510 x 765 pixels and a bag-of-features model in combination with a support vector machine, and I found no research using contrast enhancement comparatively with other pre-processing steps using CNNs, let alone on low resolution dermoscopy images. This leads to the first research question.

***Q1**: Are pre-processing steps which reduces obscuring artifact presence and enhances distinguishability that aid lesion segmentation, able to provide an increased performance to a CNN when diagnosing melanoma in low resolution images?*

Convolutional neural networks have proven to be very successful within computer vision tasks as many current state-of-the-art object recognition CNNs have shown this to be true, even outperforming humans, on datasets they were designed to classify such as the ImageNet or Cifar-10 dataset. The Cifar-10 dataset contains very low coloured resolution images(32x32), while the ImageNet data consists of millions of high resolution images that are usually resized to between 224x244 and 256x256 pixels depending on the network. The reason behind choosing these types of networks and evaluating them for this task instead of current state-of-the-art networks specifically designed for skin cancer, is because I want to be able to compare my findings to these networks related to skin cancer as well as see how general object recognition networks perform at diagnosing melanoma. The state-of-the-art CNNs used in the ImageNet competition is at the very top of computer vision as they are able to consistently and with extreme accuracy classify millions of images into thousands of different categories and show a high ability of generalization[14]. I want to find out if these types of networks at the top of object recognition in images are capable of extending beyond their point of purpose and classify melanoma effectively. CNNs have been used for skin cancer detection with promising results, and some of these CNNs have been inspired by

general image recognition models(i.e ResNet-50 or AlexNet) and then fine-tuned, optimized and extended on for the specific task of skin cancer detection. However, In this thesis, state-of-the-art object recognition models are not altered and their original architecture is preserved and evaluated on this task with images of a much lower resolution which they were not designed for. It is important to note that all models designed for other classification tasks are trained from scratch without transfer learning involved. The reason transfer learning is not used is because the huge differences in the classification tasks(for example melanoma classification compared to classifying trees/planes), difference in the image resolution, increased training time and size of the model[68].

***Q2**: How well do general object recognition CNNs perform in melanoma detection with low resolution images which they were not designed for?*

This thesis is based upon evaluating different deep learning models' performances towards the classification of melanoma in low resolution images, in order to give an overview of achievable accuracies in this task and what type of architectural network structures and pre-processing steps accumulate the best results in this regard. I will also, just like most of these networks are designed for the ImageNet dataset, design and optimize a CNN specifically towards this task.

***Q3:** Can I of design a CNN employing low resolution images and achieve equal or near-to state-of-the-art performance of existing research using CNNs with higher resolution images, for skin cancer detection?*

4. Related work

The incidence rate of melanoma worldwide continues to escalate quickly as it has been doing for the past 50 years and the main cause being exposure to ultraviolet radiation, where the risk increases drastically with prolonged or intense exposure[20][19]. For more than two decades, computer aided diagnosis of melanoma, has been an active area of research and is still being heavily invested in as its clinical applications have tremendous appeal. For the past 20 years, the potential development of automated skin cancer detection systems with high performances have been very attractive to researchers in medical image analysis and computer vision fields[19]. There is plenty of research literature available on different approaches and methods for the classification part of this thesis. Among these classifiers, the more commonly used to handle these types of classification problems include deep learning networks(CNNs)[21, 27, 57, 65, 67, 26], support vector machines[19, 65], bayesian classifiers[80], Multi-layer perceptrons[23], linear classifiers[66], decision trees[81], principal component analysis[54] and k-nearest neighbours[80][2]. Quite a bit of literature have also accumulated in the area of another challenging problem in this field that relates firmly to skin cancer detection, which is lesion segmentation. Lesion segmentation is comprised of extracting the lesional area by drawing an accurate border around it to exclude normal surrounding skin. Lesion segmentation literature cover a lot of different methods that can be implemented to tackle this problem either individually or by combining multiple techniques to achieve the best results. Some of these researched methods encompass probabilistic modelling, active contours, clustering, histogram thresholding, edge detection, graph theory and more[19].

Y. Li and L. Shen conducted research using a deep learning approach for the detection of melanomas in dermoscopic images on the ISIC 2017 testing set containing a total of 2000 images of different resolutions[21]. Three tasks were performed; Lesion segmentation, feature extraction and classification, achieving accuracies of 0.922, 0.914 and 0.852 respectively. Segmentation accuracy was measured by comparing the

network's result against a ground truth. They used a straight forward convolutional neural network for the feature extraction task, whereas the other two tasks(lesion segmentation and classification) was handled by two fully convolutional residual networks that made up a deep learning framework. Furthermore, a lesion index calculation unit was developed to cultivate the coarse results from classifying lesions through the calculation of a heat map of distance. All images of lesions was resized to a resolution of 320x320 and then used for training.

C. Barata, M. Celebi *et al[22].* investigated the effect four different color constancy algorithms had on the classification rate of dermoscopic images using a bag of features model. The four algorithms experimented with included Gray World, Shades of Gray, General Gray World and max-RGB, compared to a baseline without any color constancy algorithms applied. They found that implementing color constancy techniques improved the classification accuracy of images of multiple sources, bettering the original accuracy, specificity and sensitivity of 63.1%, 55.5% and 71% to 77.8%, 76% and 79.7% respectively. All algorithms outperformed the baseline and received very similar results with a maximum deviation in accuracy of 0.8%. The shades of gray algorithm showed the best performance, surpassing the runner-up(general gray world) by 0.2% in accuracy.

M. Sheha, M. Mabrouk *et al[23]*. presented an automated approach for detecting the presence of malignant melanomas in skin lesion images through texture analysis without the use of any segmentation methods. Their approach incorporated a gray level co-occurrence matrix for extracting features present in the skin lesions and for the classifier, they used a multilayered perceptron. Two different techniques was used with the classifier, an automatic(dividing the data into 60% for training, 20% for validation and 20% for testing) and a traditional technique(dividing the data into 75% for training and 25% for testing) , where they found the latter method to be slower, but performing better and achieving an accuracy of 92% on the test set. The data set these experiments were conducted on consisted of a total of 102 dermoscopy images of lesions, split evenly between benign melanocytic nevi and malignant melanoma, where

the classifier's training and testing data used 75% and 25% of the entire data set. All images were resized to a scale of 512x512 pixels during the pre-processing stage.
D. Cai, K. Chen *et al[25].* conducted research where they experimented with three state-of-the-art models, namely, *AlexNet, GoogleNet and VGG-Net* on images of low resolution in four categories; cars, birds, dogs and flowers. The original images where cropped and resized to a fixed resolution of 50x50. They found that these convolutional neural networks assume that the input data is of adequately high resolution and image quality as their performance quickly crumbles when presented with low resolution images. They then proposed a novel deep learning model that is resolution-aware and through a comprehensive set of experiments, found that their model performed better than the CNNs when classifying low resolution images containing fine-grained objects on a persistent basis. As an example, they tested AlexNet on the data set of cars and was able to improve the initial accuracy of 50.4% to 63.8%.
A. Esteva, B. Kuprel *et al[26]*. researched the effect the GoogleNet inception v3 convolutional neural network architecture, pre-trained on the 2014 data set from the imageNet competition, had on classifying skin cancers. Their data consisted of a total of 129,450 clinical images, whereas only 3374 of these images were dermoscopy images. Contained in this data set are 2032 unique diseases. They conducted two validation experiments for checking the performance of the classification rate of their network; The first test consisted of three prediction classes of benign lesions, malignant lesions and non-neoplastic lesions. The second validation test involved nine different classes of diseases. The results they observed of the two validation tests were 72.1% ± 0.9% and 55.4% respectively, compared against certified dermatologists performing the same tasks under the same conditions receiving a peak accuracy of 66% and 55%. The CNN was trained on labelled clinical images according to their true class and resized to a fixed resolution of 299x299 pixels. The aim of this study was to demonstrate a classification algorithm that is generalizable and they found that the performance of their CNN achieved a level of classification competence matching real world expert dermatologists.

L. Yu *et al[27]*. proposed a novel method in 2016 consisting of two main stages for melanoma recognition that takes advantage of deep convolutional neural networks. They created a convolutional residual network for segmenting lesions in images and integrated it with other existing deep CNNs for the task of classifying the lesions to make a deep learning framework for classifying melanomas. The data set they incorporated in their training and testing were acquired from the ISBI 2016 melanoma detection challenge set. They split their dataset, a total of 1250 images, into 900 images for training and 350 for testing. Each initial image was made up of a resolution of 1024x768 which was then cropped containing the segmented lesion and resized to a fixed resolution of 250x250 and then fed to the networks used for classification. Their data set was then artificially augmented using different rotational orientations of the same image, noise and translation. The best accuracy they observed was 0.855 using the DRN-50 network combined with segmentation, and 0.828 without segmentation.
T. Brinker *et al[62]*. provided a novel systematic review of skin lesion classification research, considered state-of-the-art, employing CNNs. They only included methods that used CNNs for the purpose of classifying skin lesions in their review, excluding CNNs used for lesion segmentation or feature classification/patterns. Results from state-of-the-art research were presented, and they observed a difficulty in comparing models against each other due to the authors of the research not divulging fully towards what methods was implemented in training, as well as using datasets that are not publically available.

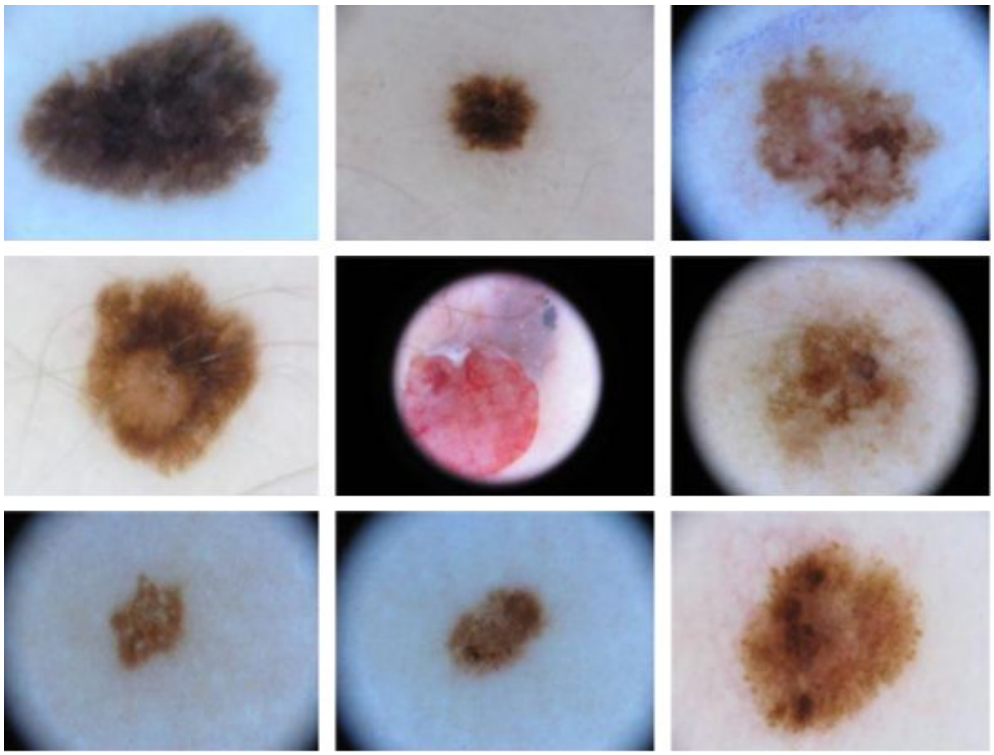
Creating a computer aided diagnosis tool for skin cancers to aid real world professionals, still is and has been, an appealing area for quite some time and as a result a lot of research in this area has been conducted, that has amassed an abundance of literature for me to absorb and learn from going forward. In my thesis, I still base my work on previous research conducted in the field, but my contribution in this field will be to give an overview of meaningful and novel comparative studies. This study includes the effects pre-processing steps exert on the classification rate of

melanoma, how well state-of-the-art general object recognition CNNs perform at this task, and if the proposed CNN, using low resolution dermoscopy images, can match state-of-the-art CNNs performance in this task. Every model in the study is evaluated with the same dataset containing low resolution dermoscopy images. To my knowledge, at the time of writing this thesis, the work I am doing is novel.

5. Data

There are few datasets containing dermoscopy images of skin lesions that are publically available. Among these, are three datasets of interest; The PH2 dataset, the ISIC archive and the HAM10000 dataset. The PH2 dataset contains a total of 200 dermoscopic images, where only 40 are classified as melanoma[5]. This size will be insufficient for the purposes of this research, but can be combined with images from other datasets. Every image in the HAM10000 dataset is available through the ISIC archive[6]. The ISIC archive, also referred to as the melanoma project, is an international collaboration and industry partnership that has been established with the intention to help reduce the mortality rate of melanoma through facilitating data gathering to be used in digital skin imaging applications[7].

Figure 1: Some sample images from the ISIC Archive of lesions



The ISIC archive contains an excess of 23000 images of all kinds of skin lesions, both malignant and benign. Out of these 23000 images, 2169 are unique melanomas and 19373 are benign. This dataset fits best towards the needs of this thesis and is the dataset I will be going with for training and testing networks to distinguish melanomas from benign lesions. Specifically, I will be using a balanced subset of the ISIC dataset. If I have the need to expand on my dataset, I have the option of combining multiple datasets or artificially augmenting it.

6. Methods

In order for me to conduct my intended research, one essential piece that needs to be acquired is adequate data that can be used in my various scientific experiments and evaluated thereafter. As discussed in the data section, I retrieved a balanced set of high quality dermoscopy images from the ISIC archive, but I also wanted to be unbiased when choosing the images in order to not choose images that would best suit my specific models and make it as easy as possible to achieve high accuracies as I wanted to train the models that are able to generalize, as there can be an incredible amount of variety and similarities between benign lesions and malignant melanomas. When gathering data from the ISIC archive, each page consists of 80 images of lesions and a user is able to select which page and which images, or all images on that page to retrieve. The data gathering approach I used was to select every image on a page that I chose in a somewhat random fashion whilst keeping track of the pages I had already retrieved images from as to avoid choosing duplicate images. After choosing a satisfactory amount of dermoscopy images, the resulting data set was made up of 2400 images intended for training and 640 images for testing, both split evenly between the two output classes. This original data set contained 3040 dermoscopy images and was created to be fed to different convolutional neural networks which requires inputs of fixed sizes. Because of the static input formats that was required by the models, all images were resized to a fixed resolution of 112x112 pixels. Cropping techniques,

specifically center cropping was not used in order to preserve possible information regarding lesions present in pictures. Normally cropping would be a great idea when sizing down to retain more pixels of the areas of interest, but I wanted to use an autonomous approach. Elaborating further, a consequence of being unbiased in selecting images to use, is that the images provided by the ISIC archive comes in all kinds of different dimensions and artifacts which can be remedied to a certain degree by resizing and employing other techniques, but most importantly the various lesional area coverages present and different levels of magnification used to take these images. Lesions may cover an entire image or the majority of it, and using center cropping techniques on these images may result in lost information in the form of important features such as artifacts/patterns present in lesions or the border shape which may be important in order to predict a solid and accurate diagnosis. Lesion segmentation is a challenging problem that deals with this that can be done manually, semi- or fully autonomously, but was not a focus in this thesis.

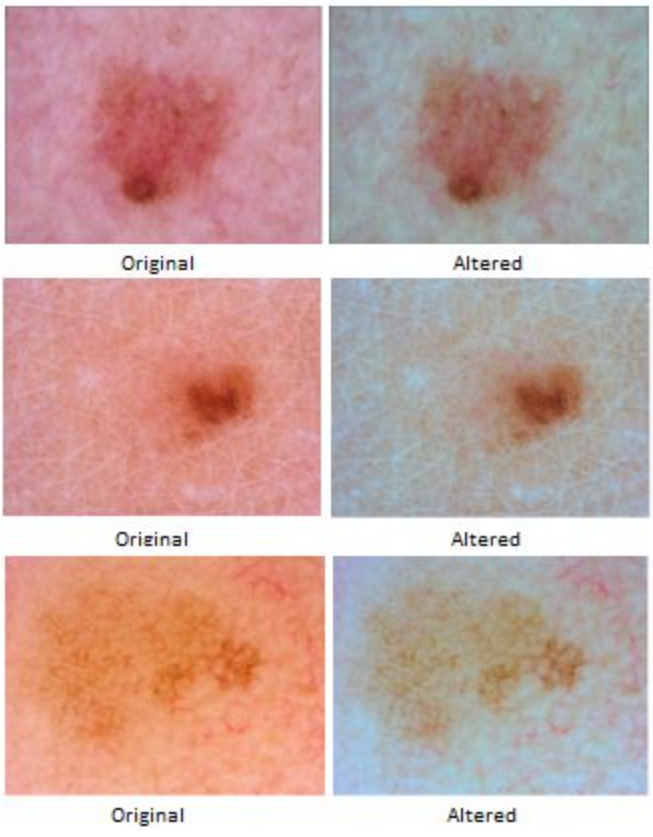
6.1 Pre-processing techniques

Employing pre-processing techniques on given raw images is an important step to facilitate lesion segmentation in order to obtain desirable results which has been a big challenge and an intriguing area of research for many years, but there are many different techniques that can be used, singularly or in combination and chosen specifically in correspondence to the data at hand. However, I wanted to evaluate models that would be able to generalize as well as possible to new images and see if these pre-processing steps, that may be an intuitively perceived improvement in the distinctive abilities for humans, contribute to improving direct classification accuracies. I chose three unique and popular pre-processing steps that can be applied to every image to remove potential artifacts and/or improve quality. Specifically, I chose and implemented popular steps that aided in reducing artifacts that may obscure parts of the lesional area and steps that enhances the distinction abilities.

Contrast enhancement

When dealing with dermoscopy images of lesions, there may be artifacts present in the image such as bubbles or strands of hair and there are algorithms that work on minimizing the effects these artifacts exert. But the most important element is to be able to distinguish the lesion from the normal skin that is not of interest surrounding it[2]. In some images, this distinction may be very hard to make or partial parts of the lesion may be occluded as it merges with the color of the background. One approach to maximizing the ability to distinguish lesion from skin can be done by appropriately enhancing the contrasts in the image to emphasize the lesional area[2]. The algorithm I am choose for this task is a white balancing method that employs the gray world assumption algorithm[15].

Figure 2: contrast enhancement applied to images.



The gray world algorithm makes the assumption that the image, more or less, contains a neutral gray color and generates an illumination estimation by taking each RGB channel in the image and computing the mean value of these channels and finally performs a normalization operation on the image using the computed mean value[15].

Enhancing the contrast in the image is especially important when creating a lesion segmentation algorithm which is a very difficult problem in itself. When segmenting the lesion the machine needs to be able to localize the region containing the lesion and this becomes much more difficult when the distinction between the lesion and surrounding skin is

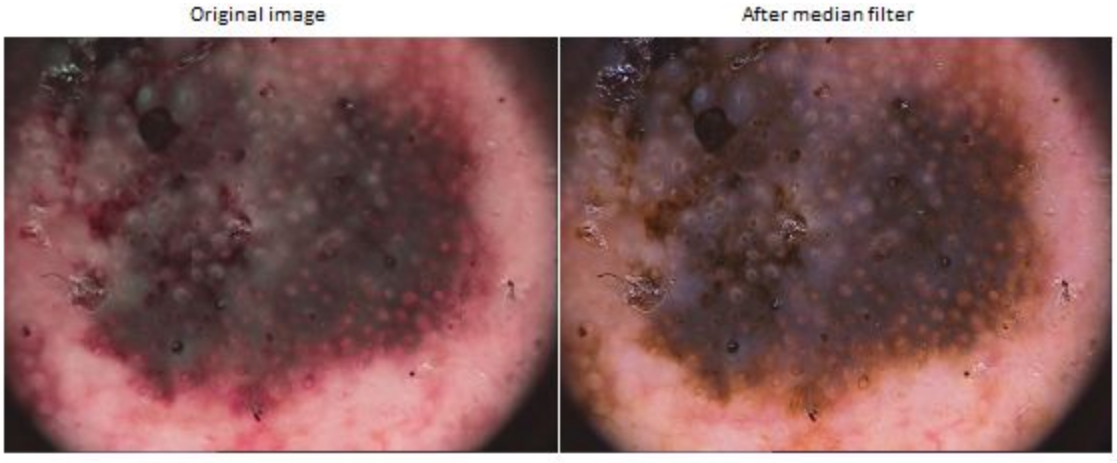
minimal. However, the idea is that this step may also be applicable when directly applying an image into a classifier as the distinctions become more apparent after this step.

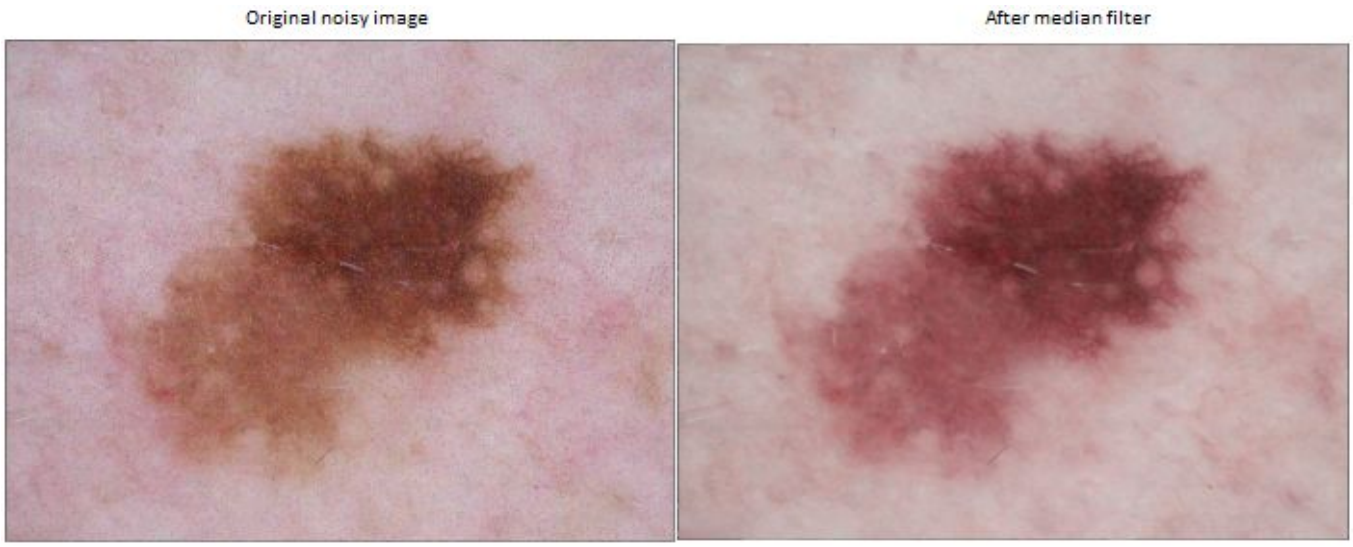
Every image in the data set was iterated through and for each individual image, a white balancing method using the gray world assumption algorithm was applied in order to enhance the contrasts in the image with the goal of making the area of interest(lesion) more easily distinguishable from the typical surrounding skin, as illustrated in figure 2. This program was created in MATLAB R2018b.

Median filter

There are many different types of artifacts that may be present in an image and one of these that may obscure parts and be detrimental to the performance of a network's prediction, is unwanted signals or noise. This step is quite a substantial one in digital image processing as it can lead to improved enhancement of an image, for instance enhancing the contrast. Noise disturbances in an image is more commonly a form of electronic noise and results from an alteration of the color information or brightness of a random nature and can originate from a number of different sources; errors in the transmission of data, camera sensors, radiation, grain effects or water droplets among other factors[28][29].

Figure 3: results of applying a median filter.





A method for suppressing unwanted noise in an image is a median filter. This is a common filtering technique that works in a non-linear fashion by running through the entries of signals individually and proceeding to use the median values of adjacent entries to replace the current signal entry[30][31]. In the first image row illustrated in figure 3, the colors change globally and smoothes the original picture giving it a more clear texture, and the last row of images shows the transformation of an image affected by salt and pepper noise using this filter. The median filter was created using Java and employs a 3 x 3 filter.

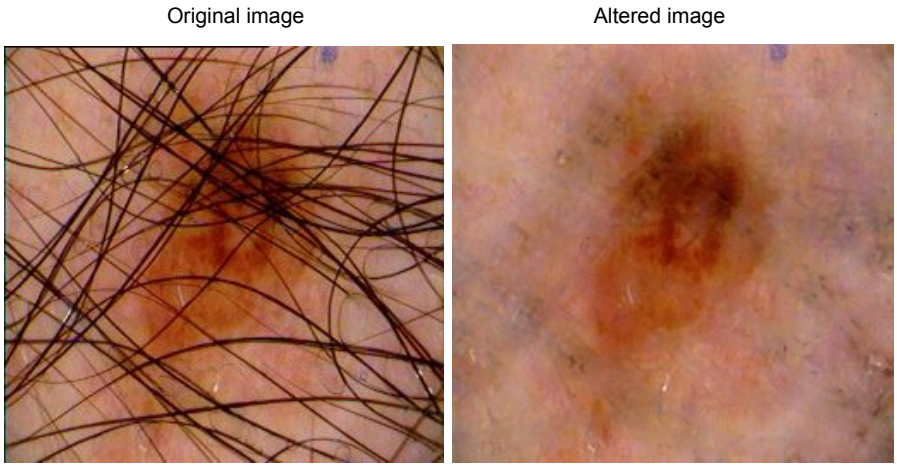
Hair removal

In recent times, an increase of studies directed at analysing melanocytic lesions for possible malignancy or benignancy using various image pre-processing techniques to alter the original image to facilitate diagnosis, such as hair removal or enhancing the contrasts to name a few, have started appearing[4]. One of the difficulties of analysing these lesions precisely however, is the possibility of occurring body hairs which may partly obscure the lesion of interest, making some of the inherent features attributed to the lesion in question hidden from view or confuse a classifier resulting in inconsistent results, in terms of segmenting the lesion from the surrounding skin or determining an

accurate diagnosis through being unable to extract all the features present that may be important to accurately determine the state of a lesion[4].

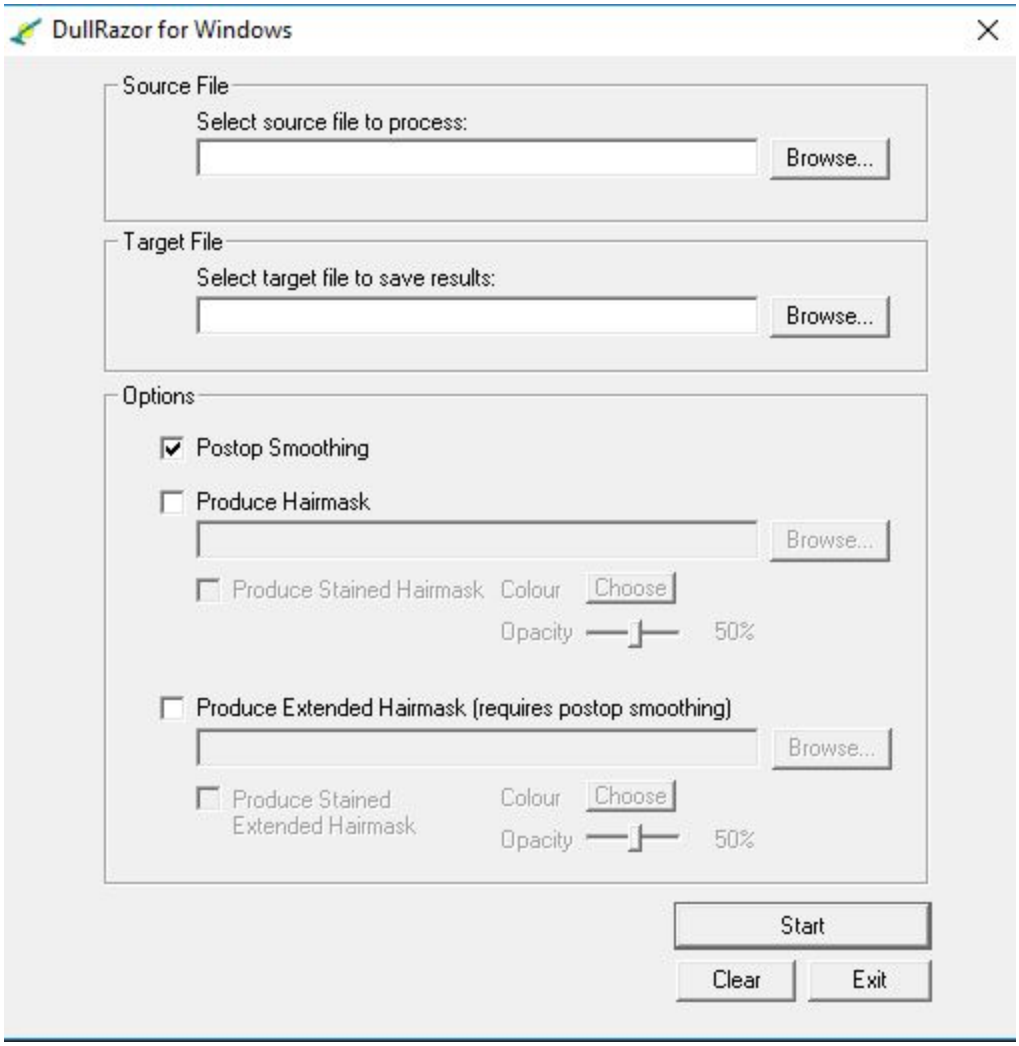
There is already an existing software program that deals with hair removal of lesions from dermoscopy images which is conveniently named DullRazor[3]. The way this program works is by first identifying the locations of the dark strands of hair pixels by using a generalized morphological grayscale closing operation[4]. The algorithm then verifies the hair pixels' shape as a long and thin structure, and finally applies a flexible median filter to smooth out the verified hair pixels that have been replaced[4].

Figure 4 from [4]:



The image displayed above in *figure 4* is the satisfactory result of applying the hair removal software DullRazor on the original image.

Figure 5: Dullrazor GUI



Above in figure 5 is a look at the graphical user interface of the hair removal software. Attempting to use this software I encountered an immensely time consuming problem. This program is only ever able to handle one file at a time as opposed to a list of files, and in order to alter an image the user is required to first input the correct path to the

image that needs to be changed, and another path for the location and new name of the altered image. The user then has to initiate the program through its start button and finally acknowledging a successful alteration of an image by hitting an ‘OK’ button that appears. Repeating this process manually when I have a dataset containing an excess of 3000 images will get extremely tedious and time consuming. I may also want to change to a different dataset to test on or extend the current dataset with images that are not artificially augmented and then this process would need to be repeated manually all over again. In order to remedy this temporal challenge I created a program in Java with the intention of automating this process. This java program works by first initiating this external hair removal software and then finding the appropriate x and y coordinates on the screen that correlates to the positions of the necessary buttons and input fields required to successfully remove hair from an image and store the altered image. These coordinates only needs to be discovered once due to this program only being run on a personal computer with a single resolution, and because the external software’s position remain unchanged after each operation. The java program then iterates over every image in a specified folder and performs the hair removal functionality of the DullRazor application on each image with a slight intentional delay to allow for DullRazor to run successfully and register the entire sequence of inputs.

This robot that I created in java managed to successfully alter 3040 images in just over an hour. This is by far, faster than the most proficient human being is capable of, saving me a lot of hours of tedious work and can also be reused if needed.

6.2 Deep learning classifiers

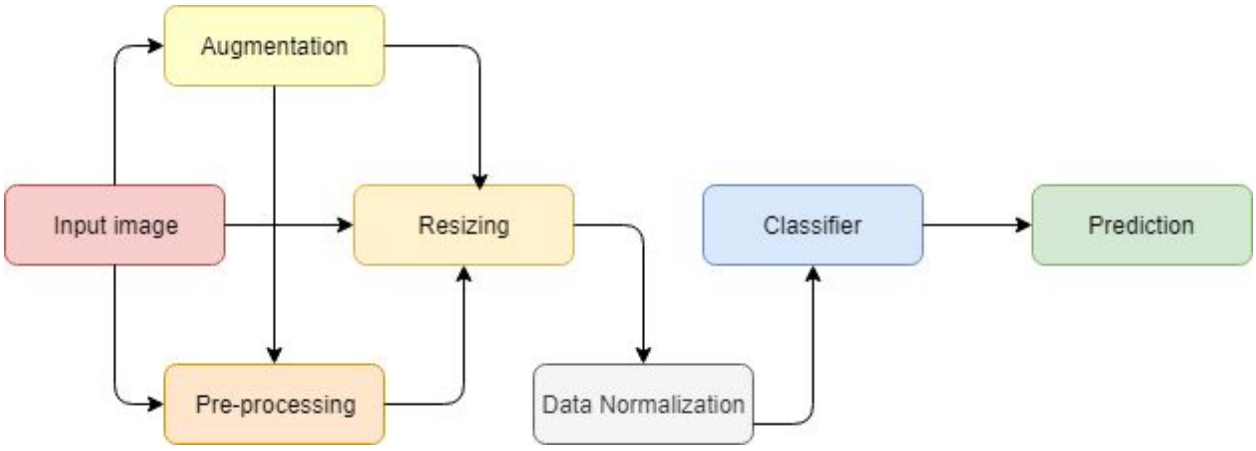
In order to predict whether an image containing a lesion was malignant or not, I used multiple state-of-the-art networks that are proven to perform extremely well on the *ImageNet Large Scale Visual Recognition Competition* and *Cifar-10* data sets. These networks include; AlexNet, ResNet-50, A model designed to classify the Cifar-10 set,

Xception and VGG-16. In addition to this I also designed a custom CNN that was inspired by a network created for the MNIST dataset[79], for the purposes of the task of this thesis. The architectures and configurations of the models will be explained in much greater detail later on in the text.

6.3 Evaluating the models

When evaluating the performance of the models I experimented with, I used metrics that are considered common practice and works well with my setup, as well as being used to compare the results I achieved to other pieces of research performances using CNNs when diagnosing skin cancers. The main metrics that is used in this thesis to measure how well the models hold up are accuracy, specificity and sensitivity which will be thoroughly defined later on. Because I am dealing with diagnosing a type of cancer, specificity and sensitivity are important metrics to consider, especially sensitivity in the case of diagnosing melanoma due to the importance of not missing an actual non-benign lesion. A confusion matrix is also used to visualize exactly how many classes were correctly and incorrectly classified which is then used to conclude the aforementioned metrics. An accuracy metric will work well for acquiring a solid evaluation of the performance of a model's predictions because of the 50/50 balance of classes that was used in the dataset[47]. As an example, in the opposite case where there is an imbalance of classes, the accuracy metric may not provide a truthful description of how the model performs in terms of accuracy, i.e 90% of class I and 10% of class II, can skew the model making it easy to receive a 90% accuracy by having the network only predicting one class[47].

Figure 6: Methodological approach



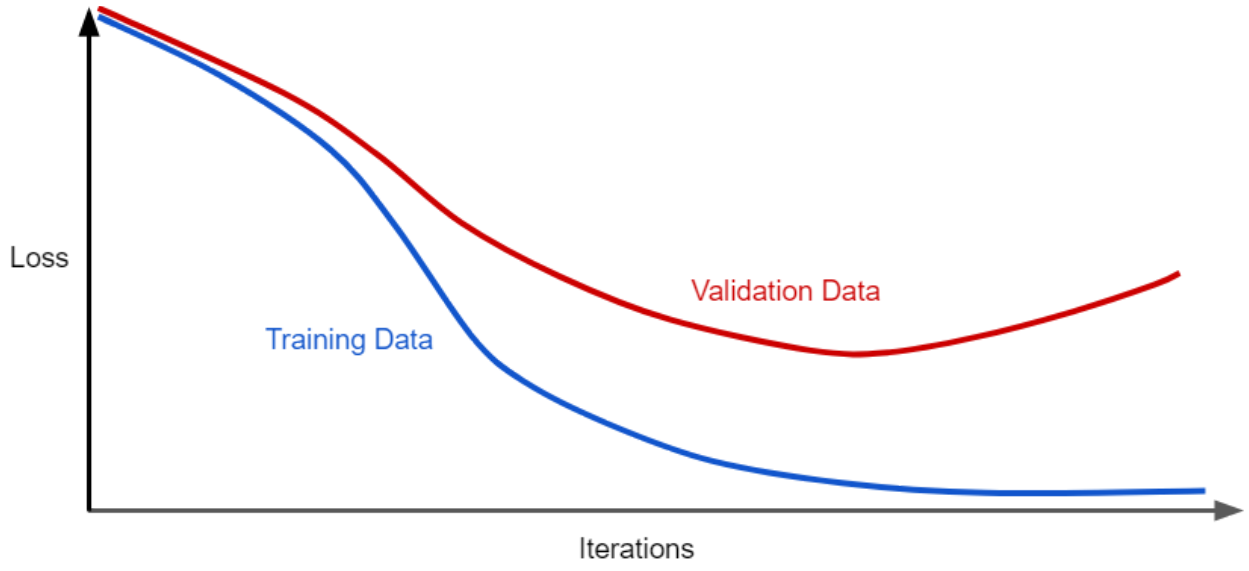
In my thesis, I experimented with different configurations of applied pre-processing steps on images with and without dataset augmentation(expanding the dataset artificially using the original images), in order to determine what configuration paired with specific models performed the best at this task. The research method I used to approach diagnosing this particular type of skin cancer is illustrated in figure 6, where an image has either been augmented through rotational variances or altered in some way by 1 or more pre-processing steps or both, before being scaled down to a predefined size that is desirable. The reason augmentation and resizing is separated from the ‘pre-processing’ item in the figure above is to give a more detailed overview of what types of configurations can be made before a prediction and evaluated, i.e both augmentation and one of the three pre-processing steps mentioned or simply a raw image that is resized before put into a classifier. This gives a better understanding of how the experiments are set up, because an experiment will not necessarily include all of these steps that could be presented as one. The next step was to normalize the input data by having its values constrained between 0 and 1. The reasons this was done was because it is generally considered a good practice in order to better the performance and stability of a neural network due to the differences in the scales of the values, i.e resulting in a network containing large weights, potentially making it very volatile

resulting in an increased error of generalization[48-49]. Furthermore, the second reason is that a gradient descent algorithm was used to train the neural networks, and the activation functions normally employed, have a range from somewhere among -1 and 1, contributing to a network with an improved performance[49]. Moving on to the next stage in the approach was to feed the prepared image to a chosen model in order to be classified, finally resulting in a prediction by the model whether the processed lesion was benign or malignant.

7. Reducing overfitting

One problem that has persisted for a long time when training a network that is important to consider and combat, is overfitting[39]. Overfitting occurs when the model fails to recognize prevailing patterns provided in the training set, and instead is able to recognizes distinct images, resulting in a poor ability to reliably generalize to new sets of images beyond the images the model was originally trained with[40].

Figure 7 retrieved from[44]



As illustrated in figure 7 above, it shows an example of a network learning from the training data as the loss function is steadily descending per iteration, however the testing/validation data does not emulate the former's accuracy, but instead, the curve turns to eventually go in the opposite direction of what is favourable. This is a case of a network overfitting and in my experiments, I use two different popular techniques to reduce the effect of overfitting.

7.1 Regularization

There are different regularization techniques available to help in remedying potential overfitting, and some of these methods are employed in the different network models I evaluate in this thesis. These regularization methods are intended to help the model to generalize better by making slight alterations to the learning algorithm powering the network to improve its performance[45]. The first regularization method that is used is called dropout. Dropout works by temporarily ignoring or removing a neuron from a specified layer in the network, coupled with every connection of that neuron(both incoming and outgoing), and this happens based on a predefined probability which is most commonly 0.5(50% chance) in the models I evaluate, and the neurons are chosen at random[42]. By using dropout, it compels the model to pick up on more robust features by diminishing learning that is done interdependently among these nodes[43]. These neurons and their inherent power become impeded after developing dependencies on each other, and dropout is proven to be effective at reducing these particular occurrences, making the network more robust overall[43]. The second method that is commonly used in these models is L2 regularization(also referred to as weight decay) and L1 regularization, which is used in some of the networks I analyze and they can both be applied to a model simultaneously. It is one of the more common types of regularization methods and it works by taking the cost function and updating this function through the addition of a new regularization term[45].

$$L1: \quad Cost\,function = Loss + \frac{\lambda}{2m} * \sum \|w\|$$

$$L2: \quad Cost\,function = Loss + \frac{\lambda}{2m} * \sum \|w\|^2$$

In L1 regularization, absolute values of a weight are penalized and may be eliminated by having their value being reduced to 0. This technique can produce models that are sparse(resulting in fewer weights) and can be useful when attempting to make a model smaller. On the other side, L2 regularization is unable to eliminate weights as their values cannot be reduced to 0, but close to it and therefore cannot produce sparse models. Instead of penalizing absolute values like in L1, L2 penalizes the square values of weights[45-46].

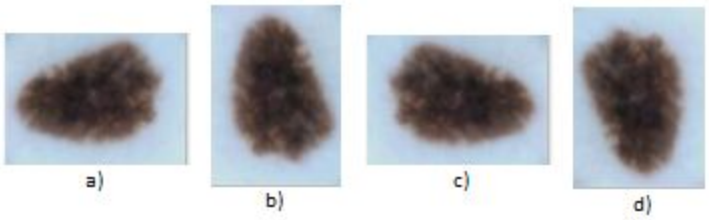
7.2 Artificial dataset augmentation

One of the most common and simpler methods to implement in order to reduce overfitting, is to enlarge the data by artificial means through transformations of the original image that preserves the corresponding label.

As discussed above in the data section, the ISIC archive contains over 19000 benign images of lesions and around 2200 images of melanomas. However, keeping a balanced training data set is essential in order to maximize the accuracy when employing traditional deep learning techniques such as neural and convolutional neural networks[8][9]. Training these traditional methods on highly imbalanced data often result in a tendency to create classifiers that overly predicts the majority class. This means that such a classifier trained on imbalanced data sets may have a high true positive rate, but a low true negative rate[8]. A study regarding this class imbalance problem was done specifically on convolutional neural networks, where they recreated the networks used on the MNIST, CIFAR-10 and imageNet datasets, and trained these

networks on imbalanced data sets. The resulting conclusion was that training these CNN's on imbalanced versions of the same dataset was harmful to the overall performance of the network[9]. A team of researchers conducted experiments with the MNIST data set using a convolutional neural network and found the best practice when preparing the data was to artificially augment it, in their case using elastic distortions as opposed to rotational variants, which contributed to improving the results substantially on the MNIST data set[41].

Figure 8: Original image a) augmented by incremental 90 degree rotations.



Although my dataset is already balanced, I want to expand it artificially to see what effect it will have on the performance of the networks in terms of accuracy and other evaluation metrics due to data being an essential part of this process, and the more data I have access to should help a CNN to perform better. The initial training set consists of 1200 malignant and 1200 benign lesional images. For each image in the initial training data set, I performed three rotations to get three extra orientations of the image in question. Three extra images were created, rotated 90, 180 and 270 degrees, from each original image as illustrated in figure 8. This increased the size of my initial training set consisting of 2400 images, with an additional 7200 images, into a new balanced training data set of 9600 training images split evenly among the two classes to be predicted, as well as increasing the size of the test set from 640 to 2560(1920 additional images). The program to artificially augment my initial dataset through different rotational orientations was created in Java using IntelliJ.

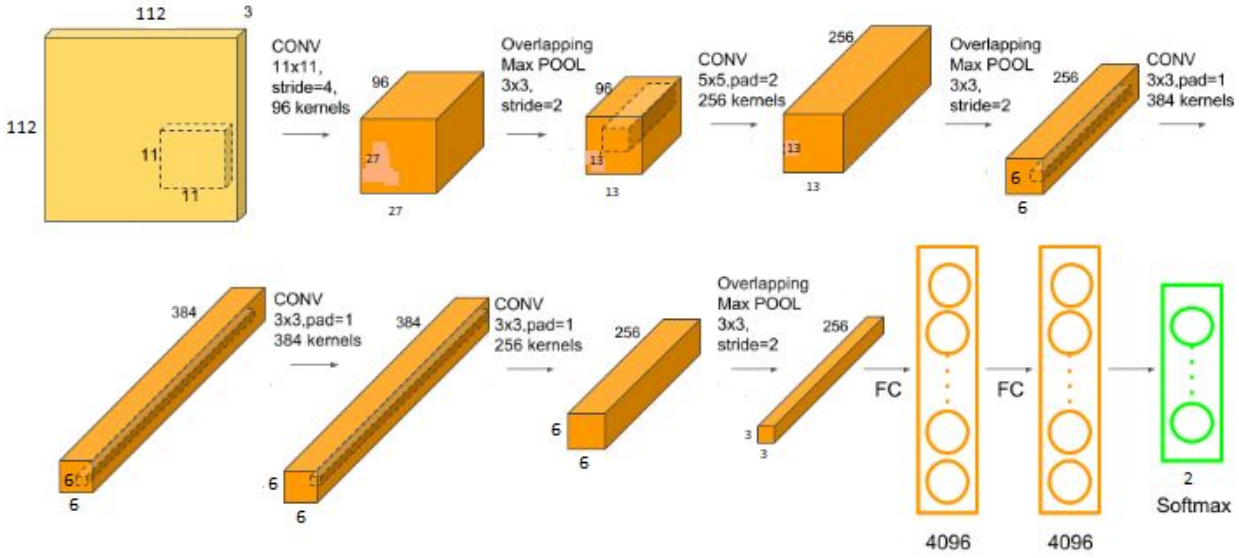
8. Deep learning network architectures

There are multiple different types of artificial neural network architectures that are used in deep learning today, and the major architectural types include recurrent neural networks, recursive neural networks, convolutional neural networks and unsupervised pretrained networks[14]. These network types have their own strengths and weaknesses and perform better on different tasks. As an example, a recurrent neural network will perform better on sequence modeling tasks such as language processing, while a convolutional neural network is best suited for image classification tasks. For the past few years, convolutional neural networks have consistently been performing at the top of image classification contests such as the imagenet competition that was held annually[14]. These CNNs are responsible for considerable advances in computer vision tasks which has immense potential for real world vision application, for instance robotics, self driven cars and medical applications[14]. Due to the great performances CNNs have obtained recently on image recognition tasks and suitability, I will mostly be focusing on evaluating the performance of different state of the art CNNs for general image recognition as well as a custom CNN model, inspired from a network created for the MNIST dataset, on my binary image classification problem.

Below, I will present the specific architectures of the different neural networks I intend to use for my research, as well as illustrations of the models which can be referred to for an easier understanding of the specific architecture of the networks. All of these networks have been interpreted and implemented in deep learning for Java based on explanations corresponding to research papers or imported from keras.

AlexNet

Figure 9 retrieved from [72]: AlexNet architecture

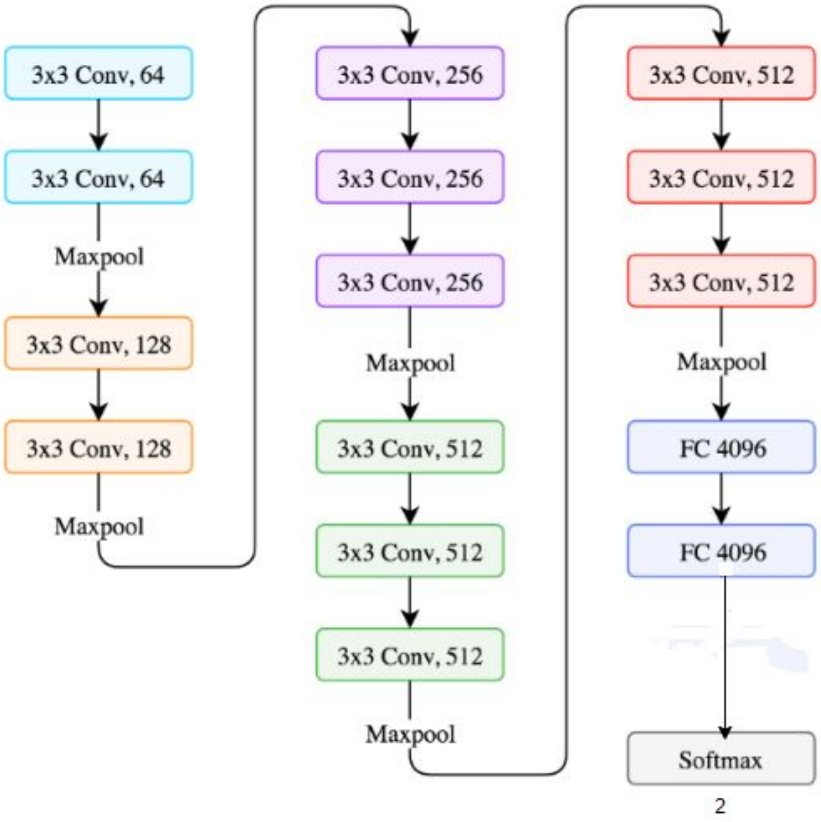


In the field of computer vision, the AlexNet paper is considered to be one of, if not the most, influential research published to date in the field[33]. AlexNet was originally designed and optimized for the purposes of classifying 1.2 million unique images of a high resolution(rescaled to 256x256) and trained on 15 million unique images from 22000 different categories in the previously annually held ImageNet competition[34]. The CNN's appearance in 2012 was considered a breakthrough and ever since, a lot of computer vision research have surfaced employing convolutional neural networks[33]. The AlexNet implementation in deep learning for java consists of 13 layers in total; five convolutional layers, two Response-normalization layers, three Max-pooling layers and three fully-connected layers. All max-pooling layers use a 3 x 3 kernel with a stride of 2 pixels. The first convolutional layer uses a 11 x 11 kernel with a stride of 4, the next convolution uses a 5 x 5 kernel, while the rest of the convolutions use a 3 x 3 kernel and

both use a stride of 2 pixels. The model employs a Nesterov's gradient descent optimizer using a learning rate of 0.01 with a momentum of 0.9. A rectified linear unit(ReLU) activation is used for the convolutional and fully-connected layers. Regularization is applied in the penultimate layer in the form of a dropout with a probability of 0.5[32]. In the final layer, the output layer, a softmax activation function is used and a negative log-likelihood is used for the loss function.

VGGNet-16

Figure 10: Architecture of VGG-16 retrieved from [73].



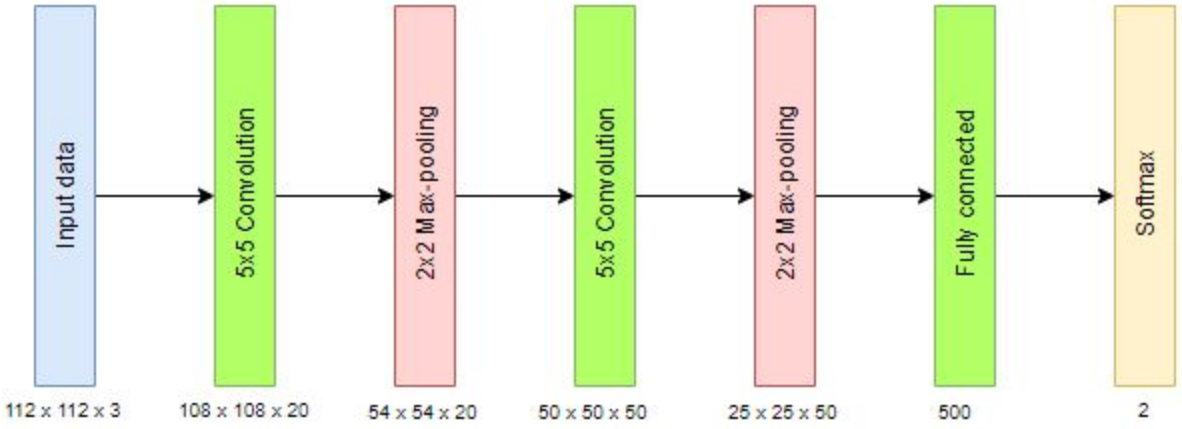
2

Another network that was designed for and submitted to the ImageNet large scale visual recognition competition in 2014 trained on 224x224 RGB images is the VGGNet[35], which is still considered state-of-the-art. The network implementation in Java is made

up of 13 convolutional layers, 5 Max-pooling layers and 3 fully-connected layers, totaling 21 layers. A filter size of 3 x 3 with a stride of 1 pixel is used for all convolutional layers, while Max-pooling layers are performed with a 2 x 2 filter and a stride of 2 pixels. This implementation employs an Adam gradient descent optimization algorithm and a ReLU activation function is applied to every convolutional and fully-connected layer in the network. A dropout regularization technique to reduce overfitting is used with a probability of 0.5 in two of the three fully-connected layers. The final layer operates with a softmax activation function and a negative log-likelihood loss function.

Proposed CNN(C-CNN)

Figure 11: Architecture of custom convolutional network

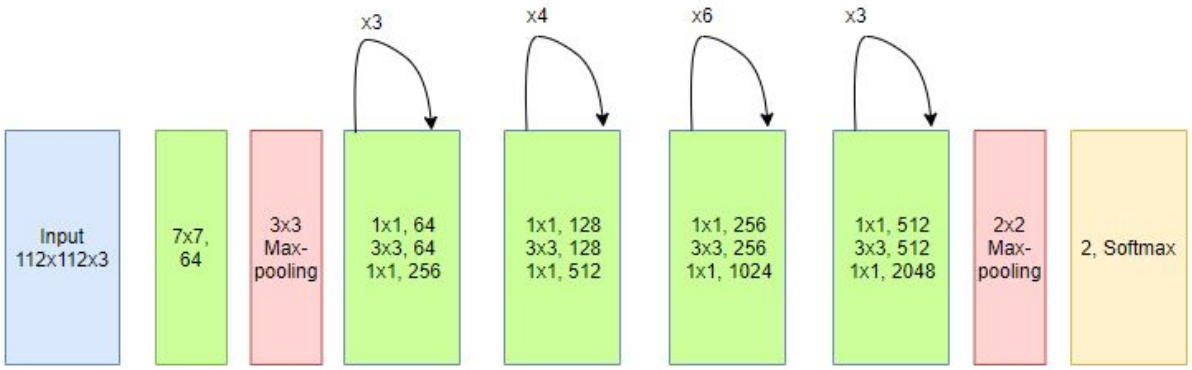


I have created a customized convolutional neural network, intended for the classification of 112x112 dermoscopy images, that has gone through a number of different configurations and different optimization approaches and the end result of these tweaks entails a model made up of 6 layers. There are two convolutional layers, and each convolutional layer is followed up by a Max-pooling layer, and then the last two layers are fully-connected. The convolutional layers in the network use a 5 x 5 kernel with a

stride of 1 pixel and the Max-pooling layers employs a 2 x 2 kernel with a stride of 2 pixels. The optimization algorithm that I have chosen is an Adam gradient descent optimizer incorporating a steady learning rate of 0.0008. Weights in the network are initialized using the Xavier algorithm that, based on the number of neurons, autonomously regulates the scale of initialization to help making sure the weights in the network are appropriately initialized. In order to address potential overfitting in my model and weight values becoming too large, I implement L2 regularization to deal with this and also apply gradient normalization in every layer of the model. In both the convolutional layers I employ Identity activation functions, but a ReLU activation is used in the penultimate layer followed by a two-way softmax activation function in combination with a negative log-likelihood loss function in the final layer.

ResNet-50

Figure 12: ResNet-50 architecture

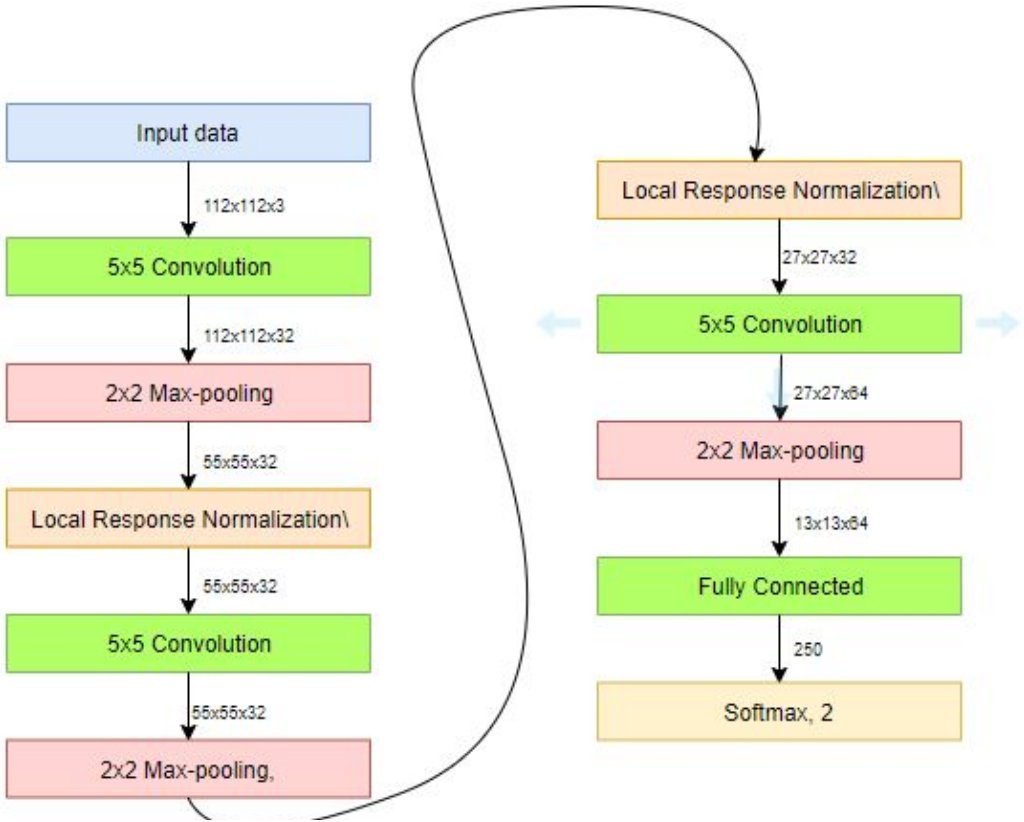


Resnet, another state-of-the-art network, was created by a team of microsoft researchers and won the ImageNet competition in 2015[34], and I will specifically be looking at a ResNet-50 implementation in deep learning for Java(DL4J) that consists of a model containing 159 layers. This is a very deep network made up of blocks of stacked layers. There are two types of building blocks that make up the body of the

network and will, for simplicity, will be referred to as block *a* and *b.* Block *a* is made up of 3 convolutional layers, each of these followed by a batch normalization layer and each batch normalization layer is followed up with a ReLU activation layer, making block *a* 9 layers. Block *b* accommodates 11 layers, 4 of these are convolutional layers, and every convolutional layer is followed up by a batch normalization layer and only 3 of these batch normalization layers precedes ReLU activation layers. A total of 12 *a* blocks and 4 *b* blocks make up the overwhelming majority of the layers in the network and is preceded by one zero padding layer, one out of the 3 stacks making up block *a* and a Max-pooling layer. The remaining penultimate and final layer, is a Max-pooling and a two-way softmax activation output layer, respectively. A negative log-likelihood function is used to calculate loss when training. The Adam gradient descent algorithm is used for optimizing the model with a learning rate of 0.1. In addition to this, an identity activation function is present in the convolutional layers of the model. Both L1 and L2 regularization is applied to the model as well as a normal distribution for the initializations of weights.

Cifar-10 Model

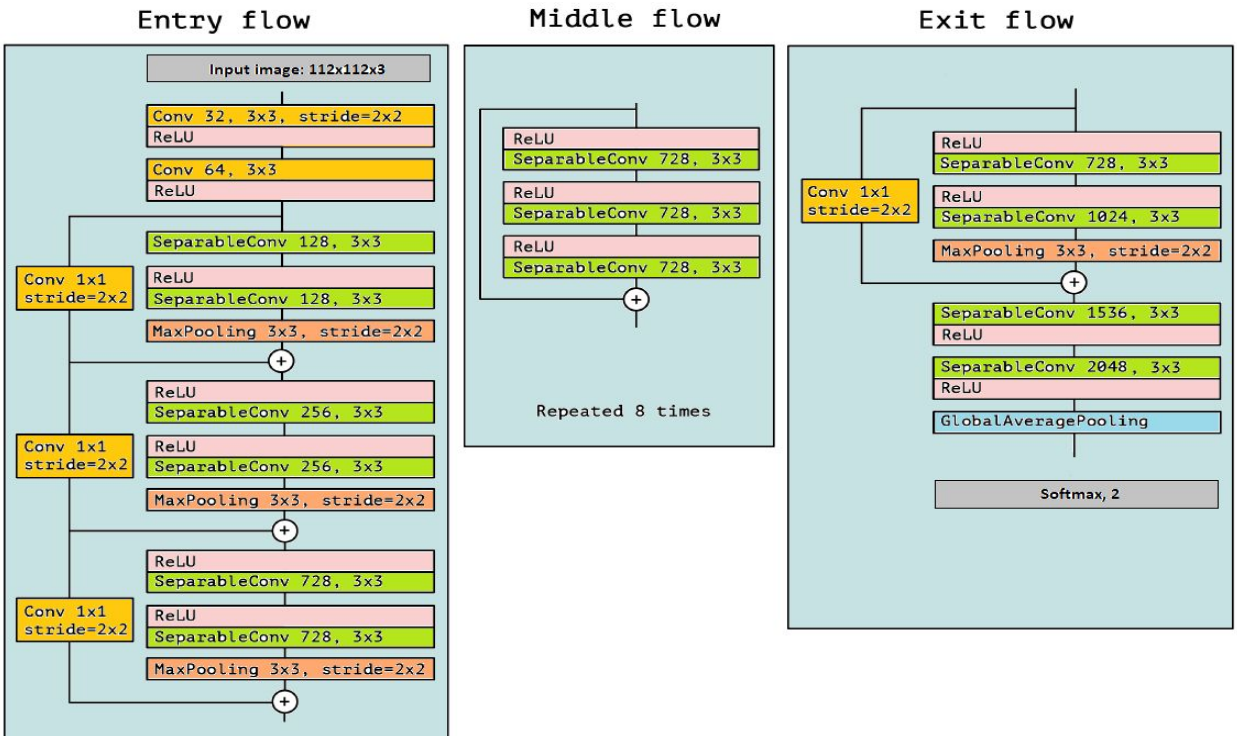
Figure 13: Cifar10-M architecture



A convolutional neural network made specifically for classifying images into the correct 10 different possible classes provided in the Cifar-10 data set[36]. This network performs well on the 60000 images contained in this data set and these image are all of low resolutions(32x32) with a colour depth of 3. Because of this, I feel it is worth evaluating this model on classifying melanomas as the Cifar-10 set shares more similarities with the data set specific to my thesis than the extensive ImageNet collection of images. This model employs a ReLU activation in every convolutional and fully-connected layers, where each convolutional layer uses a 5 x 5 kernel with a stride of 2 pixels and a padding of 2. In total, this network entails 10 layers, 3 of which are convolutional layers, each followed by a Max-pooling layer and the first two pooling layers are directly accompanied by local response normalization layers. The 9th layer is a fully-connected layer, incorporating regularization through a dropout function with a probability of 0.5 of activating. The final layer calculates the loss through a negative log-likelihood algorithm and makes use of a two-way softmax activation function. Weights in the network are initialized using the Xavier technique. In addition to a dropout function in the penultimate layer, L2 regularization is also used. In order to optimize the model, a straight-forward stochastic gradient descent algorithm is implemented.

Xception

Figure 14: Xception architecture retrieved from[74].



The Xception module[37], is a network inspired by GoogleNet's Inception architecture[38], and have been proven to marginally outperform Inception-V3(formerly known as GoogleNet) on the data set it was specifically designed for(the ImageNet data set). In this algorithm, the inception modules it was inspired from, have been changed out with depthwise separable convolutions. This novel network architecture implementation in java is made up of 58 layers. As illustrated in figure 14, the Xception model contains 6 conventional convolutional layers, 13 depthwise separable convolutional layers, 19 batch normalization layers, 14 ReLU activation layers, 4 Max-pooling layers, one global pooling layer and one fully-connected output layer.
An Adadelta gradient descent optimization algorithm, and a ReLU weight initialization is used in the network. The middle flow of this network is repeated 8 times as shown in the figure. All depthwise separable convolutional layers in the model employ kernels of size 3 x 3, while normal convolutional layers use both a 3 x 3 kernel with a stride of 2 or 1 pixel and a 1 x 1 kernel with a stride of 2, and the Max-pooling layer use a 3 x 3 kernel with a stride of 2. The network operates with L2 regularization, a two-way softmax function in the output layer with a negative log-likelihood loss function.

9. Experiments

In this section, the results from evaluating the networks will be presented using different configurations. A configuration entails a set of models trained and evaluated using a unique combination of techniques, such as applying a specific pre-processing step with or without augmentation, and comparing it against a baseline evaluation of a configuration of models trained on raw images without any pre-processing or augmentation involved. A comparison of how other networks, tackling a similar problem, performed from other existing state-of-the-art research in the field will be presented to evaluate the best model in this thesis. In accordance with the research questions, these configurations will involve evaluating the performance of all the networks previously

defined against different pre-processing steps applied to images with and without an augmented data set to reveal what contributes to improve the performance in this task. Specifically, to evaluate the performances of these deep learning networks that are trained and tested on low resolution images. Applicability of all involved state-of-the-art CNNs will be evaluated on this task under these specific conditions, as well as a comparative evaluation of the best performing model for low resolution melanoma recognition against related existing research.

All models are trained using a total of 100 epochs, where one epoch is defined as an entire iteration over the training and testing set as I evaluate the model after each epoch. Models experimented with non-augmented data sets is trained using batches of 20 images at a time, whereas the expanded data set through augmentation models use batches of 80 images at a time. These models are trained through supervised learning, where each dermoscopy image is labelled in correspondence to its true class. Delving a bit further on the structure of the data setup, there are two different main folders aptly named *training* and *testing,* and each of these folders have two sub folders, titled *0* and *1,* where *0* represents a collection of dermoscopy images of benign lesions and *1* is responsible for lesions classified as melanoma. Images retrieved from the ISIC archive come in all different sizes, and due to the networks I am using requires a fixed input, all images have been resized down from their original resolution to 112x112 pixels. The original data set is split into images, 80% for training and 20% for testing. All tables of model performances are based on results from applying the classifier on the test set. The augmented version of the data set has 9600 images for training and uses the original test set(640 images) for testing. The remaining 1920 images from the augmented test set is instead used as a validation set to provide an unbiased evaluation of the generalization ability of the best performing model and configuration to be compared to other relevant existing state-of-the-art research employing CNNs.

9.1. Evaluation metrics

When evaluating models in this thesis I will be using a variety of different evaluation metrics to assess the performances of the different deep learning models I will be training using the predefined data set extracted from the ISIC archive. The metrics I will be using are defined below and are important to understand and be able to refer back to as I will be using them frequently in my evaluations and comparisons of the models.

True positive(TP): The number of correctly classified melanoma by the model.

True negative(TN): The number of correctly classified benign lesions by the model.

False positive(FP): The number of incorrectly classified melanomas by the model.

False negative(FN): The number of incorrectly classified benign lesions by the model.

Sensitivity: The proportion of actual melanomas that are predicted to be melanomas.
TP / (TP + FN)

Specificity: The proportion of actual benign lesions which are predicted to be benign.
TN / (TN + FP)

Accuracy: The proportion of correctly classified lesions based on all predictions.
(TP + TN) / (TP + TN + FP + FN)

9.2 Evaluation of models

A lot of different permutations of configurations can be made with the experiments I conducted by combining multiple pre-processing steps with or without augmentation,

calculated to a total of 16(4 to the power of 4) possible configurations for each of the six models used. Instead of presenting all the findings, the configurations presented will be ones that provide a baseline for each unique step, and combinations of setups if they produce an improved performance. The presented results in this section will directly apply to assist in providing information and insight into the previously stated research questions. All training of the networks in this thesis have been performed on a personal computer with 16GB ram, ZOTAC GeForce GTX 1070 8GB Mini GPU, Intel(R) Core i5-7600 3.50GHz Kaby Lake processor, EVGA BQ 650W power supply and a B250m Bazooka motherboard.

Table 1: Evaluation of models using raw images as input

Model	*Accuracy*	*Sensitivity*	*Specificity*
C-CNN	0.7803	0.63125	0.9375
AlexNet	0.5671	0.968	0.1656
VGG-16	0.7891	0.65625	0.921875
ResNet-50	0.5	1.0	0.0
Cifar10-M	0.5046	0.009	1.0
Xception	0.5	0.0	1.0

In *table 1*, the performances of the different networks that are trained on purely raw images without having had any steps applied to them besides the necessary resizing is shown. The best performances in terms of accuracy is provided through the VGG-16 model(78.9%), followed closely by the customized CNN(78.4%), with both of them retaining a high level of specificity and low level of sensitivity. The rest of the networks performs poorly in comparison to this accuracy with a deviation of at least 20% in error rates. ResNet-50 did only achieve an accuracy of 50%, but a sensitivity of 100% due to

the fact the network only predicted 1 class(melanoma), identical to the Xception model with the exception of reversed sensitivity and specificity. AlexNet proved to be the third best performing model.

Table 2: Evaluation of models using preprocessed images through hair removal

Model	*Accuracy*	*Sensitivity*	*Specificity*
C-CNN	0.7140	0.5093	0.9187
AlexNet	0.4125	0.6218	0.203
VGG-16	0.7453	0.5656	0.925
ResNet-50	0.5	1.0	0.0
Cifar10-M	0.5265	0.2	0.8531
Xception	0.5	0.0	1.0

The results from applying the DullRazor hair removal software to the networks is illustrated in *table 2*, where the effect of hair removal has proven to be detrimental for the performances of *VGG-16, C-CNN* and especially AlexNet being the first model to dip beneath 50%(14.4% reduced accuracy compared to *table 1*) as they all performed better without this step being applied. Cifar10-M is the only one that benefits from this step.

Table 3: Evaluation of models with an applied median filter

Model	*Accuracy*	*Sensitivity*	*Specificity*
C-CNN	0.7703	0.6218	0.9187

AlexNet	0.503	1.0	0.006
VGG-16	0.7828	0.6375	0.928
ResNet-50	0.5	1.0	0.0
Cifar10-M	0.5046	0.009	1.0
Xception	0.5	0.0	1.0

Table 3 provides an overview of the effects of non-augmented images, that have been through median filtering, have on the performances of the various models. ResNet-50, Cifar10-M and the Xception models deliver identical results to *table 1*, and a slight decrease in Cifar-10M's accuracy and sensitivity compared to having the hairs removed instead. The last three models experienced unfavourable results in comparison to providing them with raw images without any pre-processing. C-CNN, VGG-16 and AlexNet's accuracy dropped 1%, 0.63% and 6.41% in that order. AlexNet's sensitivity improved slightly due to all melanomas being predicted correctly, as well as a drop in specificity because most of the benign lesions were also predicted to be melanoma. C-CNN and VGG-16 on the other hand show a slight decrease in the ability to predict melanoma correctly(sensitivity), the same occurring to specificity in the case of C-CNN while VGG-16's specificity increased minimally.

Table 4: Evaluation of models using preprocessed images using Gray World

Model	*Accuracy*	*Sensitivity*	*Specificity*
C-CNN	0.78125	0.6531	0.9093
AlexNet	0.575	0.25	0.896
VGG-16	0.81093	0.6968	0.925

ResNet-50	0.5	1.0	0.0
Cifar10-M	0.5015	0.006	0.9968
Xception	0.5	0.0	1.0

Contrast enhancement applied to images in *table 4* shows consistent improvements in accuracy compared to *table 1* that illustrates the effects no pre-processing steps have on the performance of models. The exception to this is ResNet-50 and Xception that remains unchanged from *table 1,2* and *table 3,* and Cifar10-M which performs slightly worse than the previously presented results. VGG-16 had the highest rate of improvement of 2.2% from its previously best performance shown in *table 1*, while C-CNN and AlexNet improved roughly by 1%. An improvement in sensitivity can be shown in both C-CNN and VGG-16 by 2.2% and 4.1% from the leading table, respectively. Specificity in the latter two models falls slightly as the sensitivity increases.

Table 5: Evaluation of models using gray world with an augmented dataset

Model	*Accuracy*	*Sensitivity*	*Specificity*
C-CNN	0.8656	0.925	0.8063
AlexNet	0.603	0.73	0.47
VGG-16	0.814	0.7593	0.8687
ResNet-50	0.5	0.0	1.0
Cifar10-M	0.6468	0.98125	0.3125
Xception	0.5	0.0	1.0

ResNet-50 and Xception were not improved by augmentation and persists in only predicting one class. *Table 5* shows the positive effect the artificially expanded data set, with contrast enhanced images using the gray world algorithm, exerts on the C-CNN, VGG-16, AlexNet and Cifar10-M models. An improved accuracy of 8.43%, a sensitivity increase of 27.2%, and a 13.12% decrease in specificity for the custom CNN is demonstrated and the same can be seen occurring to the VGG-16 model as well(0.3% accuracy and 6.2% sensitivity improvement, with a 12.2% specificity loss). The specificity loss between C-CNN and VGG-16 has a deviation of 0.92%, while the last two metrics' improved results are far superior in C-CNN's performance. Cifar10-M also benefited from this configuration with an improved accuracy of 12.03% from its previously best performance in *table 2*, as well a massive change in sensitivity and specificity from its previous results(now predicts most lesional images to be a melanoma, instead of benign). Improvements in AlexNet's performance is also observed, bettering its previous best accuracy from *table 4* by 2.8%, 48% in sensitivity and a decrease of 42.6% in specificity, mimicking the changes in performance Cifar10-M experienced in terms of sensitivity and specificity reversals.

Table 6: Introducing a validation set of 1920 images for C-CNN and VGG-16's build from *table 5*

Model	*Accuracy*	*Sensitivity*	*Specificity*
C-CNN	0.8125	0.747	0.878
VGG-16	0.7137	0.9664	0.4609

Table 5 displays the results from introducing an unbiased validation set of 1920 images to C-CNN and VGG-16 for evaluating the models' ability to generalize to new data. C-CNN outperforms VGG-16 on the validation data by roughly 10%, while suffering a 5.31% loss in accuracy from its best result in *table 5*, as well as a 17.8% decrease in sensitivity paired with an increase in specificity by 7.17%. Comparing VGG-16 to its previously best performance on the test set to *table 6*, a decline of 10.03% accuracy

and 40.78% specificity is observed, and an increase in sensitivity by 20.71%. Overall, a more balanced distribution in sensitivity and specificity is achieved by the first entry.

Table 7: Comparison of my best result(C-CNN) against other existing research

Model	*Accuracy*	*Resolution*	*Dataset size*	*Classifier*
C-CNN	**81.25%**	**112x112**	**12160**	**CNN**
Esteva et al[26]	72.1%	299x299	129450	CNN
Li et al[21]	85.2%	320x320	10976	CNN
J. Kawahara et al[67]	79.5%	227x227/ 454x45	1300	CNN
A. Lopez et al[65]	81.33%	224x224	496	CNN
Esfahani et al[57]	81%	188x188	6120	CNN
Yu et al[58]	83.51%	224x224	724	CNN
T. Pham et al[75]	89%	299x299	6762	CNN

The results displayed in *table 7* are only indicative of being in the same proximity in terms of accuracy, and not a precise and reliable performance comparison to existing research due to differences in the data and task. Every CNN in *table 7* use a unique dataset, some with or without dermoscopy images and a few classify more than two lesion classes.

In *table 7 Cargani et al's[63].* research employing a CNN that received 89.2% accuracy was not displayed in this table due to the systematic review by *T. Brinker et al[62]*, not finding this performance plausible because of a lack of ability to discuss the results and the readability of their approach. Because this thesis is based upon CNNs, *table 7*

specifically includes other state-of-the-art research that employs CNNs as the classification method. However, there are other state-of-the-art methods that are used for classification that achieve similar, and considerably better performances that are excluded from this table, such as principal component analysis[54], Multi-layer perceptrons[23], K-nearest neighbours[64], linear classifiers[66] and support vector machines[19, 65]. All the CNN classifiers demonstrated in the table above do not deal with a binary classification problem as some of them attempt to classify multiple type of skin cancers and not only malignant melanomas.

In *table 7* my best performing, and best generalizable model(C-CNN) is compared to state-of-the-art research within the field focused on utilizing convolutional neural networks as classifiers. C-CNN operates with the lowest resolution out of all the models and incorporates a data set size that is only surpassed by *Esteva et al's* dataset[26]. The latter research used 129450 images, 3374 of which were dermatoscopic images, and a three-way classification(Benign, Malignant and non-neoplastic lesions) using 757 different classes for training, as opposed to this binary task. C-CNN is shown to surpass 3 out of 7 models in accuracy. *Esfahani et al's* dataset contained only non-dermoscopic images of lesions[]57, and *Kawahara et al* trained their network using 10 labelled classes and an equal classification output[67]. The remaining proposed methods exceed C-CNN's performance.

10. Discussion

It is important to mention once again that a majority of the state-of-the-art deep learning networks used in this thesis to diagnose melanoma have originally been designed and evaluated on the ImageNet dataset using a specific resolution(224x224 - 256x256) and the Cifar-10 dataset(32x32). They were not specifically created for a binary classification problem, but rather to recognize and classify objects into thousands of different classes such as cars, cats, dogs, busses and much more. All of the classes they categorize have distinct inherent features which makes them more distinguishable

from other classes in most cases, for example a plane versus a tree. However this is not always the case with a malignant melanoma versus a benign lesion. Distinguishing a malignant melanoma from a harmless lesion is a substantially more difficult and complicated task due to the similarities they have in common and different variations, which correlates understandably towards why medical professionals in the field only achieve a diagnostic specificity of 59% due to uncertainty when distinguishing melanoma from atypical lesions[11, 61]. When diagnosing melanoma, a standardized set of guidelines are usually followed known as the ABCDE rule[50-51]. This involves checking a melanoma for an asymmetrical shape, irregular border, the coloration(if there is multiple colors present or unusual color distribution), diameter and the evolution of the lesion over the span of time, the latter being the most important factor in deciding whether it is a melanoma besides an excision. However, back to the point of the difficulty of distinction, a benign lesion may possess features normally attributed to a malignant melanoma such as an irregular border, asymmetrical shape or unusual coloration, and in turn a melanoma may possess a symmetrical shape and/or smooth border. Especially in the latter case when a melanoma is in its early stages, as it has not yet been allowed to evolve beyond a very small point which is when it is almost always curable. This is why this is a very challenging task to do with great precision in comparison to the previously mentioned image recognition tasks and should be kept in mind while going forward with discussing the results of the models tested.

Illustrated in the tables in the result section, the ResNet-50 and Xception CNNs were never able to improve beyond a 0.5 classification rate due to always predicting one class. These two models, regardless of the configurations of both pre-processing steps and dataset size provided, were never able to learn from the data. The assumption I am making as to the reason for the failure in learning is divided into two parts. The first takes into consideration the differences of resolution the models were designed for(256x256) and what was used to train them(112x112). The models assume inputs of certain quality and resolution and begins to falter when these requirements are not

met[25]. The second important bit of information to consider is the different structures of the models. Both of these models are similar in the regard that they are both are considerably deep, with ResNet-50 and Xception having 159 and 59(in addition to 6 layers which are repeated 8 times) layers, respectively. These models seem to be too complex to accumulate knowledge of patterns from low resolution lesional images. I mentioned what methods were used to reduce overfitting earlier in this paper and I find it worthwhile to mention that lowering a model's complexity can help in this regard, although it is dependant on the data provided and on the other hand, the depth or complexity of a network may result in overfitting(the dataset fits too well)[52]. I make the assumption that the depth contributes to the model becoming too complex for the data I am using to experiment with, resulting in ResNet-50 and Xception being unable to generalize to new images provided as well as the inability to learn due to heavy fluctuations in training loss. The difference in resolution is also taken into account for the the models' performance paired with their complexity.

The VGG-16 model greatly surpassed most of the models evaluated in almost every configuration setup used, and its performance in comparison to the rest that were originally designed for classifying higher resolution images or lower in the case of the CIFAR dataset, was very reasonable. However, the VGG-16 model seemed to show signs of overfitting, an inability to generalize well to the testing set, and particularly the validation set, as the difference in training and testing error rates are relatively high[55]. The idea with the loss function is to gradually decrease as training progresses, to indicate that the model is improving by calculating the deviations of predicted classes in correspondence with their true label at each training iteration[56]. In the case of the VGG-16 network, the training loss descends to an extremely low number, more or less close to 0(0.00005%), but the predictions for the testing set has an error rate deviation of around 18-22% from the training accuracy, across all of the experiments leading to an observation that the model tends to overfit with the data it is provided. VGG-16 employs dropout techniques in two of its last three layers, and the model performs reasonably well at this task considering the task's complexity, but overfitting persists.

Additionally, my assumption is also correlated to the enormous variety in possible permutations of lesions which is what makes this such a challenging task. Overall, VGG-16 performed surprisingly well considering the task and resolution it was provided, and introducing additional augmentation techniques/data could have contributed to reduce overfitting(this applies to all models).

AlexNet originally employed a Nesterov's gradient optimization algorithm with a learning rate of 0.01 and a momentum of 0.9, however this resulted in an inability to learn and make predictions that weren't solely based upon a single class. With this learning rate, AlexNet started off only predicting one single class in the test set in the first 30 epochs, before showing signs of learning between epochs 31 and 37, but then reverted back to its prediction rate of the earlier epochs throughout the remaining training phase. Fine tuning this hyperparameter(learning rate) from 0.01 to 0.1 made a big difference. After this alteration, AlexNet started to produce unique test results consistently with some fluctuations in both training and testing performance. The training loss tended to oscillate, but the general training loss trend was decreasing with these fluctuations. In the augmented configuration Alexnet was unable to consistently learn and improve with a batch size of 80 and followed the pattern previously described. Reverting back to a batch size of 20 that was used for all models without an augmented data set, made the necessary difference for this model to start learning. Observations of apparent overfitting was clear as the loss in the different configurations ended between 0.05 to 0.25 with a peak testing accuracy of 60.3% in *table 5*. Although augmentation was used in combination with AlexNet's implementation of l2 regularization and 0.5 dropout in the third- and second-to-last layers, a problem with overfitting was still clearly apparent.
D. Cai et al[25], conducted experiments on a few state-of-the-art networks, as explained earlier, such as AlexNet with low resolution images(cars, bird etc) as opposed to high resolution, and found that the performance of AlexNet dropped significantly when provided low resolution images. It is sensible that, when resizing down, important details and/or information may be lost as pixels are removed. As an example, one of the

highest dimensions of the original images in my data set was resized down from dimensions upwards of 4500 * 5000 * 3 (totaling 67.5 million pixels) to 112 * 112 * 3(37,632 pixels), resulting in a loss of information in pixels. In essence, although substantially more computationally expensive, it is more favourable for the model to train with images of a higher resolution in order to accumulate more information about patterns, specific features and more as there are more pixels to retrieve that type of information from.

Cifar10-M displayed clear signs of oscillation in every configuration, with a heavy bias towards predicting lesions to be benign, with the exception of *table 5* where the bias was transferred to predicting melanoma when augmentation was introduced. Constant fluctuations in testing accuracy was observed with deviations varying between 0.1% to 14% and more importantly training loss never went below 0.68. Not being able to learn may be a cause of using too few epochs(100), or the architecture of the model does not fit well with the data it is provided(originally designed for 32x32 images). Consequently, a reliable and consistent performance is not achieved using this model.

Designing and optimizing the C-CNN required a lot trial and error with network parameters to find the optimal performance for the data I provide. I observed that the filter size, padding, strides and outputs used in the convolutional and max-pooling layers illustrated in *figure 13,* proved to deliver the best performance. Increasing or decreasing these parameters had a pernicious effect on accuracy, and this also extended to the output in the fully-connected layer(500). Initializing weights using the Xavier algorithm performed the best to all others that were tested, which included normal distribution, Identity, Uniform, Relu, 0, 1 and Sigmoid initialization of weights. The activation functions that were tested encompassed Sigmoid, Softmax, Identity, Relu, Tanh, Selu and Leakyrelu. The best result was achieved using identity activations in the convolutional layers, relu in the fifth dense layer(fully-connected) and softmax in the output layer, in that specific order. Using Adam to optimize the gradient descent

outperformed Nesterov and RmsProp using a learning rate of 0.0008(anything else performed worse).
I found the Negative log likelihood loss calculations to be among the most common loss function used, but for the sake of certainty I experimented with cross entropy, mean squared error, mean absolute error, poisson and hinge loss to no avail. Introducing gradient normalization in the layers contributed to a better performance, but adding L1 regularization in combination with L2 regularization did not have a positive effect. Employing dropout at different probabilities(0.1% to 0.8%) resulted in a noticeable exacerbation of the performance. I also experimented with adding additional layers to increase the complexity of the model such as layers containing batch normalization, convolutions, max-pooling, local response normalization and additional dense layers, but I was unable to obtain a better performance than using the parameters displayed in *figure 13.* It is difficult to say why an increment of 0.001 in the learning rate contributes positively or negatively to the performance, as it is with a lot of the other parameters I experiment with. Looking at kernel size, strides and padding, they all contribute to deciding the size of the outputted filter in their respective layers that the following layer will incorporate. This portrays that the outputted filters and channel output and input in the layers fit well with the image resolution that is used(demonstrated in *figure 13*).It should be said that I have not attempted every permutation of parameters as this would not be feasible, and there may exist combinations of parameters in the model that would increase performance while maintaining the same amount of layers.

In the result section, *table 1*, provides a baseline to compare the evaluated models with single pre-processing steps in order to verify whether an improvement is achieved using a particular step or not. Hair removal using DullRazor had an interesting and unexpected effect on the models that had a change in performance from the latter table. This step proved to have a detrimental effect on C-CNN, VGG-16 and AlexNet as their accuracies dropped by 6.63%, 4.38% and 15.46% respectively. All these 3 models, with the exception of VGG-16 experienced a decrease in sensitivity performance, while only

C-CNN's specificity was decreased in this configuration. Intuitively, I expected this to improve performances on classification as it does with lesion segmentation, although that is a different problem. As it tends to improve segmentation(although this is a different problem where only drawing a precise border around the lesion is the goal) due to removing hairs that may obscure certain areas or overlap lesion borders, it may be correlated to the resolution of images. The hair removal software replaces pixels by using bilinear interpolation and continues to "restore" or smooth out the pixels of hair that were previously removed by applying a median filter that is adaptive. These removed hairs leave a trail in in the image that is visible to the human eye as illustrated in *figure 1*, due to the actual pixels being removed are replaced by the average weight of the four surrounding pixels which may also be a possibility for the loss in accuracy, although I would argue that it should still be an improvement out of intuition as it is performed on all images, although the network sees it differently. A very interesting article brings up a third possibility: the effect hairs growing from a lesion have on the diagnosis. *A. Bodemer M.D[69]* states that hairy moles can be very reassuring because cancerous lesions are unusual structures, and hair growth in a natural bodily mechanism. Elaborating further, he states that hair follicles beneath healthy moles results in hair growing through the lesion, but if the lesional area is damaged(in this case from the development of melanoma), hair follicles may be harmed resulting in a halted growth. A few other pieces of online material substantiate this hypothesis[70-71]. Interestingly, he also states that he has not yet in his career, observed a cancerous lesion with hair growing through it. However, this is not a sure-fire way of diagnosis and hair follicles may grow on normal skin surrounding a cancerous area and when they reach a certain height they may dip across and hover above lesions, ultimately being included in dermoscopic images without directly growing through lesions. A review of the dataset revealed a majority of hairs present in images of an estimated three to one ratio in favour of the class containing benign lesions. This may be the most likely cause towards why these three models performed better on images containing hairs as they are more prevalent in benign cases.

Median filtering as a pre-processing step had a similar effect on the models as hair removal, but with a smaller decrease/increase in accuracy compared to the baseline configuration. The exception was that Cifar10-M's performance was identical. The smaller decrease in accuracy for VGG-16, C-CNN and AlexNet from its baseline configuration may be attributed to hair still being present after a median filter is applied. And similarly to hair removal which also uses a type of filtering to smooth the removed hair pixels, only applying a median filter received a better performance when the hair was preserved which may be related. This is a beneficial step to use when removing noise from an image, preserving the edges of the lesion as it is quite popular in various lesion segmentation algorithms, but was not beneficial in most models for the case of using it directly for classification. A potential reason for this is the median filter replacing pixel values by a calculated median of neighbouring pixels and smoothing the image, consequently resulting in an alteration of the original image, although this step is conducted on the entire dataset. On a second note, it may be very dependant on the data set as my data set did not contain noticeable noise from a quick visual inspection of the data.

The final pre-processing step involving color constancy/contrast enhancement, was the only one to have a positive effect on the accuracy of the majority of the models, with the exception being a negligible decrease in Cifar10-M's performance. C-CNN had a tradeoff between sensitivity and specificity in favour of the former, while VGG-16 saw an improvement in both areas. Intuitively, a human would gain from enhancing contrasts in dermoscopy images as the distinctions between lesional and cutaneous areas become more apparent and identifying the extent of the lesion is emphasized, and from my findings, it seems that the best contending networks also benefited from this step. This step, when augmentation was introduced, provided the best performance of all models(with the exception of ResNet-50 and Xception which remained unchanged).

The last research question in this thesis deals with how close the best performance from the models I train to diagnose malignant melanoma, compares to other computer aided assistance classification research that deals with recognizing skin cancer in dermoscopy images using CNNs. All of the models I compare my best result against, use images of a considerably higher resolution to train and test with, smaller data sets(apart from one) and may include lesion segmentation as a part of their approach. The best achieved accuracy I received in my experiments were 81.25%, with a sensitivity of 74.7% and specificity of 87.8%. This result came, to my own surprise from a CNN that was originally designed and tested on the MNIST data set with an accuracy of 99%, which I extended upon and optimized through a long series of trial and error in order to find out what would contribute to improving the performance and what was detrimental to the classification rate. This customly optimized CNN is of a relatively low complexity(6 layers), but as shown in table 5 in the result section it held up, even outperforming quite a few of the other proposed methods used in published research articles employing CNNs, but as described earlier this does not indicate that this approach is better as the datasets are different. In *table* 7, most of the CNNs I compare to and outperform, use a smaller data set such as the case with *Kawahara et al's* proposed method[67]. They only used 1300 images for both training and testing(none of which were dermoscopic images), and I had an excess of 12000 images to train and test with at my disposal, which made a considerable difference to the original 3040 images before augmentation. An improvement of 0.25 in accuracy was shown against *Esfahani et al's approach[57],* but it is important to restate that they used 10 different output classes as opposed to two, a smaller dataset(50% the size of the one used in this thesis) and the lowest resolution of 188x188 of all, only beaten by my data. The final CNN that was outperformed was *Esteva et al's* research[26], but they used a three-way accuracy, where non-neoplastic lesions were additionally predicted, as well 757 unique training classes split on among all training data. This differs from the approach presented in this thesis, and they only had access to 3374 dermoscopic, consequently resulting in an unfair comparison as the scope of my results encompasses

a much smaller piece and constraints around the two tasks differ. The best performing model, *pham et al[75],* is not a very reliable interpretation of the network's performance using an accuracy evaluation metric as their dataset was highly imbalanced, containing only 18.08% Melanoma, while the rest were benign. This bias can be seen in their sensitivity and specificity where the model leans heavily towards predicting benign lesions(55.6% and 97.1% respectively).

Without the expansion of my data set, my best results on the test set would have been the performance of the VGG-16 model, trained on images with externally enhanced contrasts.

The general architecture of CNNs are not a new discovery and has been around for a while, but for the past two decades researchers have had better, and more access to data, graphical processing units, as well as a significant increase in the computing power of machines[59]. Due to augmentation, I have been able to minimize the effect of overfitting on the proposed custom CNN and this has played a significant role in the diagnostic abilities on the test set from the next closest configuration in terms of performance, translated into a 8.43% and 27.19% increase in accuracy and sensitivity, respectively. However, I was very skeptical receiving this type of accuracy and did not find it plausible and followed up by introducing a validation set roughly 20% of the size of the training set and the accuracy and sensitivity decreased to 81.25% and 74.7%, while specificity rose to 87.8%. This validation set definitely gave a better indication towards a more reliable performance of the model's ability to generalize to new data, although luck may still be in play, but this is difficult to reliably answer. The C-CNN's accuracy on the test set in *table 5* tended to experience oscillation, and started to stabilize more with a smaller deviation in loss fluctuations after around 70 epochs. Throughout the training process, the test accuracy tended to slightly outperform the training accuracy across all configurations, and the validation accuracy had a very similar performance(very slightly higher) to that of the training results.

The custom convolutional neural network I used in this binary classification task, limited to a very specific low resolution and without any form of cropping or lesion segmentation, showed some very promising results given the constraints around it. As discussed earlier, the intended idea of a computer aided system would not be to replace real world professionals, but rather to help them in their decision making as an auxiliary tool. An interesting example, is Google's augmented reality microscope for detecting cancer[60]. Google researchers created and applied a prototype deep learning tool to the real world problem of detecting breast cancer in lymph nodes and prostate cancer, and received very promising results. Essentially, the process involves the user looking into a microscope on a targeted image, and this prototype updates the user through the optical field in real time by presenting the analytical results of the machine learning algorithms used. This prototype was created to assist pathologists in their diagnosis and the responsible team believes that this tool has a potential for a significant impact in health globally. Keeping that in mind, when comparing my best result to the average diagnostic statistics of pathologists, the model presented in this thesis has a slightly higher sensitivity, and a considerably higher specificity. This is a good base performance where improvements and optimizations are definitely a possibility, and it is feasible that an auxiliary deep learning tool or as a separate function, much similar to what Google presented in 2018, could help to assist pathologists in making better decisions towards a diagnosis. Good techniques to help aid professionals to identify melanoma, especially in its early evolutionary stages when it is almost always curable and is more difficult to diagnose before it has been allowed to develop further, is an immense field of interest as it can contribute to saving lives and cut down on unnecessary procedures.

11. Conclusion

In this thesis I have investigated CNNs' performance on melanoma detection using low resolution images with accuracy, sensitivity and specificity metrics. Out three evaluation metrics, sensitivity is especially important to consider when dealing with melanoma detection. When there is uncertainty in an expert's evaluation regarding a cutaneous lesion examination through using dermatoscopic imaging which is the prevailing standard of care, the next logical step is to perform a biopsy in order to either confirm or disprove cancerous activity. This procedure involves removing a piece of the tissue or cells of interest and then sending the sample to a lab for analysis and testing[10]. This step is currently a necessity for making definitively sure if a sample is cancerous or not and may be a costly procedure in some areas of the world[10]. As it currently stands, manual professional atypical skin lesion diagnosis through analysis of dermatoscopic imaging, achieves at best a sensitivity of 90%, however the specificity is considerably lower at 59%[11, 61].This is where an artificially intelligent run computer aided diagnostics tool comes into the picture due to the tremendous recent growth of computational power and ability to outperform humans in image classification tasks such as the dataset and state of the art algorithms used in the imagenet competition. However, in order for such a system, particularly in the case of melanoma diagnosis, to be accepted and see a widespread implementation globally in the medical field, it has to be exceptionally accurate and most importantly avoid false negatives. The ideal of this type of a computer aided system is to be able to assist professionals in their field, such as dermatologists and pathologists, in a positive way when classifying skin cancer and not to replace them. As stated previously in this text, melanoma is almost always treatable if detected early and misdiagnosing melanoma as a false positive results in an unnecessary surgical procedure, whereas misdiagnosing melanoma as a false negative(benign lesion when it is actually cancerous) has a detrimental effect on the survival rate as the disease is allowed to go undetected and develop further. Due to the

dangerous effects it can have of diagnosing false negatives, this type of misdiagnosis is far more imperative to reduce than its counterpart.
VGG-16 and Alexnet experienced problems with overfitting, but were the only networks apart from the proposed custom CNN to gradually show signs of learning when training at a consistent pace, while the rest(Cifar10-M, ResNet-50 and Xception) showed signs of heavy oscillation both in training and testing and an inability to learn as the general training loss trend did not consistently decrease. As for the deep ResNet-50 and Xception models, they were originally designed for 256x256 images and when provided a considerably lower resolution, were unable to learn by always predicted one class. The assumption I made was attributed to the depth, as they were noticeably deeper and more complex than the other ImageNet models which performed better. Cifar10-M showed, similarly to the latter two models, heavy signs of fluctuations, prediction imbalance(heavily favouring one class) and a difficulty to learn.

In this thesis, it has been shown that employing pre-processing steps to dermoscopy images for classification has a noticeable impact on the performance of networks. Hair removal and median filtering proved overall to have a negative effect compared to the baseline, while using gray world algorithm to enhance contrasts to emphasize the distinction between lesion and skin, affected the performance favourably. Adopting artificial augmentation to expand the original dataset, in combination with the latter pre-processing step, proved to be the configuration that demonstrated the best performance. The best performing network was observed to be the proposed CNN with an accuracy of 81.25%, a sensitivity of 74.7% and a specificity of 87.8%. The runner-up(VGG-16) showed a considerable bias towards predicting lesions to be malignant(96.64% sensitivity with 71.37% accuracy) resulting in a very low specificity, but consistently outperformed the rest of the state-of-the-art image recognition models(designed for ImageNet and Cifar-10) used in this thesis by a sizeable margin. The proposed model was shown to generalize well to new data and achieve near-to state-of-the-art research performances diagnosing skin cancer using CNNs for

classification, by using images of a considerably lower resolution. This is only indicative of receiving accuracy in the vicinity of existing state-of-the-art research that the proposed CNN was compared to, due to all models using unique datasets for evaluation. However, it has been shown that low resolution dermoscopy images are able to achieve promising results using CNNs. A higher resolution is certainly preferable due to more pixel information being retained, that may give a better indication of features or patterns present in lesions for the classifier to accumulate better knowledge from. The proposed CNN did not match the sensitivity that the average dermatologist have on atypical lesions, and misclassified roughly 25% of all malignant lesions as benign which would be unacceptable when lives are at stake. In this regard, the validation results from VGG-16 is preferred as it would reduce the risk of actual melanomas being misdiagnosed and allowed to grow, albeit a consequent of this is a lot of false positives(unnecessary excisions). The performances demonstrated in this thesis have shown great promise, but would not be adequate for real world professionals to rely on at this time. Future work that could aid in improving this proposed model, could involve lesion segmentation, more data, additional pre-processing steps and metadata(age of the person, location of lesion and such). In this thesis, using images of a lower resolution, I have shown the ability to emulate accuracies of state-of-the-art CNNs diagnosing skin cancer using images of a higher resolution by applying augmentation and image pre-processing.

12. Rerefences

1. The skin cancer foundation(2019): What is Melanoma?
Available at: https://www.skincancer.org/skin-cancer-information/melanoma
Accessed in February, 2019.

2. N. Mishra, M. Celebi (2016, January 27): An overview of melanoma detection in dermoscopy images using image processing and machine learning.
Available at: https://arxiv.org/ftp/arxiv/papers/1601/1601.07843.pdf. Accessed in August, 2018.

3. Lee T, Ng V, Gallagher R, Coldman A, McLean D. DullRazor(1997, November 27): A software approach to hair removal from images. Computers in Biology and Medicine 533-543

4. T. Lee(1997, April 7): DullRazor download site.
Available at: http://www.dermweb.com/dull_razor/.
Accessed in November, 2018.

5. T. Mendonca, P. Ferreira, J. Marques, A. Marcal, J. Rozeira(2013, July 3-7): PH2 - A dermoscopic image database for research and benchmarking.
Available at: https://www.fc.up.pt/addi/ph2%20database.html*. Accessed in January, 2019.*

6. P. Tschandl, C. Rosendahl, H. Kittler(2018, August 14): The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6091241/
Accessed at 12 march, 2019.

7.The ISIC Archive(2016): Overview. Available at:
https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/isicArchive
Accessed in march, 2019.

8. Q. Wei, R. Dunbrack(2013, July 9): The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics.
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706434/
Accessed at 15 march, 2019.

9. M. Buda, A. Maki, M. Mazurowski(2018, October). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks.
Available at: https://arxiv.org/pdf/1710.05381.pdf. Accessed at 17 march, 2019.

10: B. Krans(2017, April 26): Biopsy.
Available at: https://www.healthline.com/health/biopsy
Accessed at 17 march, 2019.

11: A. Bhattacharya, A. Young, A. Wong, S. Stalling, M. Wei, D. Hadley(2017, July 26): Precision Diagnosis Of Melanoma And Other Skin Lesions From Digital Images.
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543387/
Accessed in march, 2019.

12: S. Tadiparthi, S. Panchani, A. Iqbal(2008, May): Biopsy for Malignant Melanoma - Are We Following the Guidelines?
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2647196/. Accessed in march, 2019.

13. Dermoscopedia contributors, A. Lallas(2018, July 25): False Negative Diagnosis.

Available at: https://dermoscopedia.org/False_Negative_Diagnosis. Accessed 30 march, 2019.

14. A. Gibson, J. Patterson(2017, July 28): Deep learning.
Available at:
https://www.oreilly.com/library/view/deep-learning/9781491924570/ch04.html. Accessed at 19 march, 2019.

15. Pi9404(2012, December 30): Color Constancy Algorithms.
Available at:
https://www.scribd.com/document/117587710/A-detailed-descriptions-and-results-for-different-color-constancy-algorithms#fullscreen&from_embed. Accessed at 19 march, 2019.

16. D. Gershgorn(2017, July 26): The data that transformed AI research-and possibly the world.
Available at:
https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/. Accessed at 21 march, 2019.

17. Wikipedia contributors(2019, March 12): ImageNet.
Available at: https://en.wikipedia.org/wiki/ImageNet. Accessed at 16 march, 2019.
18. D. Ciresan, U. Meier, J. Masci, J. Schmidhuber (2012, January 23): Multi-column deep neural network for traffic sign classification.
Available at: http://people.idsia.ch/~juergen/nn2012traffic.pdf. Accessed at 22 march, 2019.

19. S. Bakheet(2017, January 1): An SVM Framework for Malignant Melanoma Detection Based on Optimized HOG Features.

Available at: www.mdpi.com/2079-3197/5/1/4/pdf*. Accessed in august, 2018.*

20. N. Matthews, W. Li, A. Qureshi, M. Weinstock, E. Cho(2017, December 21): Chapter 1: Epidemiology of Melanoma.
Available at: https://www.ncbi.nlm.nih.gov/books/NBK481862/. Accessed at 20 march, 2019.

21. Y. Li, L. Shen(2018, February 11): Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network.
Available at: https://arxiv.org/ftp/arxiv/papers/1703/1703.00577.pdf. Accessed in August, 2018.

22. C. Barata, M. Celebi, Senior Member, IEEE, J. Marques(2015, May 3): Improving Dermoscopy Image Classification Using Color Constancy.
Available at: http://vislab.isr.ist.utl.pt/wp-content/uploads/2016/02/15_JBHI_2.pdf. Accessed at 20 march, 2019.

23. M. Sheha, M. Mabrouk, A. Sharawy(2012, March): Automatic Detection of Melanoma Skin Cancer using Texture Analysis. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.2922&rep=rep1&type=pdf Accessed at 20 march, 2019.

24. M. Elgamal(2013): Automatic Skin Cancer Images Classification.
Available at: https://thesai.org/Downloads/Volume4No3/Paper_42-Automatic_Skin_Cancer_Images_Classification.pdf. Accessed at 20 march, 2019.

25. D. Cai, K. Chen, Y. Qian, J. Kämäräinen(2017, March 15)
Available at: https://arxiv.org/pdf/1703.05393.pdf. Accessed at 25 march, 2019.

26. A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, S. Thrun(2017): Dermatologist-level classification of skin cancer with deep neural networks. Available at: https://www.nature.com/articles/nature21056.epdf?author_access_token=8oxIcYWf5UNrNpHsUHd2StRgN0jAjWel9jnR3ZoTv0NXpMHRAJy8Qn10ys2O4tuPakXos4UhQAFZ750CsBNMMsISFHIKinKDMKjShCpHIlYPYUHhNzkn6pSnOCt0Ftf6. Accessed at 30 march, 2019.

27. L. Yu, H. Chen, Q. Dou, J. Qin, P. Heng(2016, December 21): Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. Available at: https://www.researchgate.net/publication/311805330_Automated_Melanoma_Recognition_in_Dermoscopy_Images_via_Very_Deep_Residual_Networks Accessed at 31 march, 2019.

28. The University of Edinburgh School of Physics and Astronomy(2007, August 20): Topic 5: Noise in Images. Available at: https://www2.ph.ed.ac.uk/~wjh/teaching/dia/documents/noise.pdf Accessed at 3 april, 2019.

29. Wikipedia contributors(2018, December 18): Image noise. Available at: https://en.wikipedia.org/wiki/Image_noise Accessed at 3 april, 2019.

30. R. Paranjape(2009): Fundamental Enhancement Techniques. Available at: https://www.sciencedirect.com/topics/engineering/median-filtering Accessed at 3 april, 2019.

31. Wikipedia contributors(2019, March 26): Median filter.
Available at: https://en.wikipedia.org/wiki/Median_filter
Accessed at 3 april, 2019.

32. A. Krizhevsky, I. Sutskever, G. Hinton(2017, May 24): ImageNet Classification with Deep Convolutional Neural Networks. Available at: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
Accessed at 7 April, 2019.

33. Wikipedia contributors(2019, March 14): AlexNet. Available at: https://en.wikipedia.org/wiki/AlexNet#cite_note-:0-5
Accessed at 7 April, 2019.

34. K. He, X. Zhang, S. Ren, J. Sun(2015, December 10): Deep Residual Learning for Image Recognition. Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf
Accessed at: 8 April, 2019.

35. K. Simonyan, A. Zisserman(2015): Very Deep Convolutional Networks for Large-Scale Image Recognition. Available at: https://arxiv.org/pdf/1409.1556.pdf
Accessed at 9 April, 2019.

36. A. Krizhevsky(2009, April 8): Learning Multiple Layers of Features from Tiny Images. Available at: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
Accessed at 12 April, 2019.

37. F. Chollet(2016, October 7): Xception: Deep Learning with Depthwise Separable Convolutions. Available at:
http://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf
Accessed at 16 April, 2019.

38. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich: Going Deeper with Convolutions. Available at:
https://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf
Accessed at 16 April, 2019.

39. Wikipedia contributors(2019, February 23): Overfitting.
Available at: https://en.wikipedia.org/wiki/Overfitting
Accessed at 19 April, 2019.

40. R. Ruizendaal(2017, May 12): Deep Learning #3: More on CNNs and Handling Overfitting. Available at:
https://towardsdatascience.com/deep-learning-3-more-on-cnns-handling-overfitting-2bd5d99abe5d
Accessed at 19 April, 2019.

41. P. Simard, D. Steinkraus, J. Platt(2003, August 6): Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. Available at:
http://cognitivemedium.com/assets/rmnist/Simard.pdf
Accessed at 19 April, 2019.

42. J. Brownlee(2018, December 3): A Gentle Introduction to Dropout for Regularizing Deep Neural Networks. Available at:
https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/

Accessed at 20 April, 2019.

43. A. Budhiraja(2016, December 15): Dropout in (Deep) Machine Learning.
Available at:
https://medium.com/@amarbudhiraja/https-medium-com-amarbudhiraja-learning-less-to-learn-better-dropout-in-deep-machine-learning-74334da4bfc5
Accessed at 20 April, 2019.

44. Google inc.(2019, March 5): Regularization for Simplicity: L2 Regularization.
Available at:
https://developers.google.com/machine-learning/crash-course/regularization-for-simplicity/l2-regularization
Accessed at 20 April, 2019.

45. S. Jain(2018, April 19): An Overview of Regularization Techniques in Deep Learning(with Python code).
Available at:
https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/
Accessed at 22 April, 2019.

46. S. Deviant(2010): Regularization: Simple Definition, L1 and L2 Penalties.
Available at: https://www.statisticshowto.datasciencecentral.com/regularization/
Accessed at 25 April, 2019.

47. A. Mishra(2018, February 18): Metrics to Evaluate your Machine Learning Algorithm.

Available at:
https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
Accessed at 27 April, 2019.

48. J. Brownlee(2019, February 4): How to Improve Neural Network Stability and Modeling Performance With Data Scaling.
Available at:
https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/
Accessed at 1 May, 2019.

49. Deeplearning4j, Skymind(2018): Preparing Data for Learning and Prediction.
Available at: https://deeplearning4j.org/docs/latest/deeplearning4j-concepts
Accessed at 1 May, 2019.

50. J. Jensen, B. Elewski(2015, February 15): The ABCDEF Rule: Combining the"ABCDE Rule" and the "Ugly Duckling Sign" in an Effort to Improve Patient Self-Screening Examinations.
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4345927/
Accesset at 5 May, 2019.

51. The Melanoma Research Foundation(2019): The ABCDEs of Melanoma.
Available at:
https://www.melanoma.org/understand-melanoma/diagnosing-melanoma/detection-screening/abcdes-melanoma
Accessed at 5 May, 2019.

52. P, Ecarlat(2017, April 28): CNN - Do we need to go deeper?

Available at:
https://medium.com/finc-engineering/cnn-do-we-need-to-go-deeper-afe1041e263e
Accessed at 8 May, 2019.

53. Q. Abbas, M. Celebi, I. Garcia(2011, October): Hair removal methods: A comparative study for dermoscopy images.
Available at:
https://perso.telecom-paristech.fr/ytendero/ima_206_project_papers/Abbas_celebi_garcia_Hair_removal_methods:_A%20comparative_study_for_dermoscopy_images.pdf
Accessed 11 May, 2019.

54. M. Ramezani, A. Karimian, P. Moallem(2014): Automatic detection of Malignant Melanoma using Macroscopic Images.
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236807/
Accessed at 11 May, 2019.

55. W. Koehrsen(2018, January 28): Overfitting vs. Underfitting: A Complete Example.
Available at:
https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765
Accessed at 14 May, 2019.

56. J. Brownlee(2019, January 28): Loss and Loss Functions for Training Deep Learning Neural Networks.
Available at:
https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/
Accessed at 14 May, 2019.

57. E. Nasr-Esfahani, S. Samavi, N. Karimi, S. Soroushmehr, M. Jafari, K. Ward, K. Najarian(2016, August): Melanoma Detection by Analysis of Clinical Images Using Convolutional Neural Network.
Available at: http://www.dental-revue.ru/Other/2003-11-02/nasresfahani2016.pdf
Accessed at 15 May, 2019.

58. C. Yu, S. Yang, W. Kim, J. Jung, K. Chung, S. Lee, B. Oh(2018, March 7): Acral melanoma detection using a convolutional neural network for dermoscopy images.
Available at:https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193321
Accessed at 16 May, 2019.

59. R. Demush(2019, February 27): A Brief History of Computer Vision(and Convolutional Neural Networks).
Available at:
https://hackernoon.com/a-brief-history-of-computer-vision-and-convolutional-neural-networks-8fe8aacc79f3
Accessed at 18 May, 2019.

60. M. Stumpe, C. Mermel(2018, April 16): An Augmented Reality Microscope for Cancer Detection. Available at:
https://ai.googleblog.com/2018/04/an-augmented-reality-microscope.html
Accessed at 19 May, 2019.

61. H. Kittler, H. Pehamberger, K. Wolff, M. Binder(2002, March): Diagnostic accuracy of dermoscopy. Available at:
https://www.sciencedirect.com/science/article/abs/pii/S1470204502006794
Accessed at 23 May, 2019.

62. T. Brinker, A. Hekler, J. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. Enk, C. Kalle(2018, October 17): Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review.
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6231861/
Accessed at 23 May, 2019.

63. P. Carcagni, A. Cuna, C. Distante(2018, July 26): A dense CNN approach for skin lesion classification. Available at: https://arxiv.org/pdf/1807.06416.pdf
Accessed at 23 May, 2019.

64. V. Pomponiu, H. Nejati, N. Cheung(2016, September): Deepmole: Deep neural networks for skin mole lesion classification. Available at:
https://ieeexplore.ieee.org/abstract/document/7532834
Accessed at 23 May, 2019.

65. A. Lopez, X. Giro-i-Nieto, J. Burdick, O. Marques(2017, February): Skin lesion classification from dermoscopic images using deep learning techniques.
Available at: https://www.triage.com/publications/cRomero-Lopez.pdf
Accessed at 23 May, 2019.

66. J. Kawahara, A. BenTaieb, G. Hamaraneh(2016, April): Deep Features to Classify Skin Lesions.
Available at: https://www.triage.com/publications/isbi2016b.pdf
Accessed at 23 May, 2019.

67. J. Kawahara, G. Hamarneh(2016, October 17): Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers.
Available at: http://www.cs.sfu.ca/~hamarneh/ecopy/miccai_mlmi2016a.pdf
Accessed at 23 May, 2019.

68. J. Janzen(2018, January 14): CNN: Transfer learning vs build from scratch.
Available at: https://joshjanzen.com/cnn-transfer-learning-vs-build-from-scratch/
Accessed at 23 May, 2019.

69. A. Bodemer M.D(2016, August 10): Ask a Doc: Are Hairy Moles Cancerous?
Available at: https://www.menshealth.com/health/a19526196/hairy-mole-skin-cancer/
Accessed at 24 May, 2019.

70. C. Cobb(2018, July 17): Is a Hairy Mole a Sign of Cancer?
Available at: https://www.healthline.com/health/hairy-mole
Accessed at 24 May, 2019.

71. I. Paules-Bronet(2019): Hairy Moles: What It Means If You Have a Dark Strand Growing Out Of One Of Your Moles.
Available at: https://www.littlethings.com/are-hairy-moles-normal
Accessed at 24 May, 2019.

72. S. Nayak(2018, May 22): Number of Parameters and Tensor Sizes in a Convolutional Neural Network (CNN).
Available at:
https://www.learnopencv.com/number-of-parameters-and-tensor-sizes-in-convolutional-neural-network/
Accessed at 25 May, 2019.

73. A. Dertat(2017, November 8): Applied Deep Learning - Part 4: Convolutional Neural Networks.

Available at:
https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2
Accessed at 26 May, 2019.

74. S. Tsang(2018, September 25): Review: Xception - With Depthwise Separable Convolution, Better Than Inception-V3 (Image Classification).
Available at:
https://towardsdatascience.com/review-xception-with-depthwise-separable-convolution-better-than-inception-v3-image-dc967dd42568
Accessed at 26 May, 2019.

75. T. Pham, C. Luong, M. Visani, V. Hoang(2018, January): Deep CNN and Data Augmentation for Skin Lesion Classification.
Available at:
https://www.researchgate.net/publication/323161914_Deep_CNN_and_Data_Augmentation_for_Skin_Lesion_Classification
Accessed at 27 May, 2019.

76. Q. Abbas, M. Celebi, I. Fondón, M. Rashid(2011, February): Lesion border detection in dermoscopy images using dynamic programming.
Available at:
https://www.researchgate.net/publication/49747582_Lesion_border_detection_in_dermoscopy_images_using_dynamic_programming
Accessed at 28 May, 2019.

77. A. Nammalwar, B. Ghita, C. Whelan(1999, January): Segmentation of Skin Cancer Images.
Available at: https://www.sciencedirect.com/science/article/pii/S0262885698000912

Accessed at 28 May, 2019.

78. R. Sumithra, M. Suhil, D. Guru(2015, March): Segmentation and Classification of Skin Lesions for Disease Diagnosis.
Available at: https://www.sciencedirect.com/science/article/pii/S1877050915003269
Accessed at 28 May, 2019.

79. Y. LeCun, C. Cortes, C. Burges(1998): The MNIST database of handwritten digits.
Available at: http://yann.lecun.com/exdb/mnist/
Accessed at 29 May, 2019.

80. L. Li, Q. Zhang, Y. Ding, H. Jiang, B. Thiers, J. Wang(2014, October 13): Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system.
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4204387/
Accessed at 30 May, 2019.

81. A. Victor, M. Ghalib(2017, March 1): Automatic Detection and Classification of Skin Cancer. Available at:
https://pdfs.semanticscholar.org/4707/f552e2083c6a1797ad23ab621aac883b8a96.pdf
Accessed at 30 May, 2019.