DESIGN, IMPLEMENTATION AND EVALUATION OF A DEEP LEARNING PROTOTYPE TO CLASSIFY NON-PIGMENTED MALIGNANT SKIN CANCER FROM DERMATOSCOPIC IMAGES

MARIA DEL PILAR AGUILERA MANZANERA

LUND INSTITUTE OF TECHNOLOGY
Lund University

Centre for Mathematical Sciences
Mathematics

LUND UNIVERSITY

MASTER THESIS

Design, implementation and evaluation of a deep learning prototype to classify non-pigmented malignant skin cancer from dermatoscopic images

Author

María del Pilar
Aguilera Manzanera

Supervisor

Ida Ardvisson

Co-Supervisor

Anna Gummeson

Examiner

Kalle Åström

June 1, 2022

Abstract

The current trends for most fair-skinned populations are that the incidence of melanoma and non-pigmented skin lesions is growing, and this growing trend will continue for the upcoming years. The emergence of deep learning networks and their promising results in solving real-world healthcare problems and improving diagnostic accuracy opens new possibilities.

This thesis consists of the creation of a preliminary deep learning network to classify non-pigmented skin lesions: Basal cell carcinoma, actinic keratosis, and squamous cell carcinoma. This network could be used to provide feedback to the dermatologist regarding the diagnosis of a lesion at Skånes University Hospital in Lund.

We started studying publicly available data sets that could be used to reach our goal. Once we had the data sets that would be used, we proceeded to train the different networks. The networks were trained using transfer learning technology, in which we used existing pre-trained model architectures to train our model. The project was developed in Python using the Keras library that runs under Tensorflow. The results for each of the experiments were compared in terms of performance, and those that obtained the best results were selected. Additionally, we studied the versatility of the models to be used in other data sets that differed from the one used for training, and compared them in terms of accuracy and bias towards certain classes. Finally, the Grad-CAM algorithm was implemented to visualise the hot spot areas on which the model based its predictions for each of the lesions.

The final conclusions of the project show promising results that open the possibility of a future real-world implementation of using a deep learning network in a clinic.

Keywords: Melanoma; Skin cancer; Dermatoscopy; Image classification; Machine learning; Artificial intelligence; Convolutional neural networks; Dermatology; Squamous cell carcinoma; Basal cell carcinoma; Actinic keratosis; Computer-aided Diagnostics; Digital dermatology

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
AK	**A**ctinic **K**eratoses
BCC	**B**asal **C**ell **C**arcinoma
CNNs	**C**onvolution **N**eural **N**etworks
DEJ	**D**ermal **E**pidermal **J**unction
DL	**D**eep **L**earning
GPU	**G**raphics **P**rocessing **U**nit
ILSVRC	**I**mageNet **L**arge **S**cale **V**isual **R**ecognition Challenge
ML	**M**achine **L**earning
NPD	**N**on **P**olarized **D**ermoscopes
PD	**P**olarized **D**ermoscopes
SCC	**S**quamous **C**ell **C**arcinoma
SSH	**S**ecure **SH**ell

A master thesis project is not only a personal effort, but requires the professional and personal support of a number of people to whom I would like to dedicate a few lines of thanks.

To my mother for being the strongest person I know and always wanting the best for us, to my father for always being there, and to my sister for telling me I am a freak and that she doesn't understand anything.

To Ida and Anna for hosting me in their apartment and our weekly meetings in which they have helped me throughout the project with advice and guiding me through the whole process. Thank you for teaching me so much.

To Åsa and the entire dermatology department at Lund Hospital for taking care of me and making me feel welcome during the 3 weeks I spent there.

To Pepe for always being there, supporting me and believing in me throughout the process. To my friends who in these two years in a new country have accompanied me as a big family, in particular to Livius, Nate, Sergio, Juan and Iker. For always listening to me when I had a problem and sharing our victories as a big team.

Chapter 1

Introduction

1.1 Context

In most fair-skinned populations, such as Australia, Norway, Denmark and the Netherlands [5], the incidence rates of melanoma and non-melanoma skin cancers are rising. In particular, in Sweden, non-melanoma skin cancer is one of the top five most common forms of cancer in 2020 by the WHO [6]. The incidence of this type of cancer is estimated to continue to increase in the next 20 years, in particular, non-melanoma types are predicted to have a higher increase with a clear difference between men and women, Figure 1.1. Additional efforts are necessary in primary and secondary prevention to contain and possibly reverse these trends. Healthcare systems face an increasing pressure to meet the demands of constraint budgets, as well as the economic impacts of skin cancer on society. The impact is influenced by both direct and indirect costs; direct costs include resources related to the treatment of skin cancer, while indirect costs refer to productivity costs associated with illness and premature mortality.

1.2 Motivation

With skin cancer being one of the cancers that continues to increase globally, the demands for this type of diagnosis are also growing accordingly. The most common types of skin cancer are Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC), the last possibly starting with Actinic Keratosis (AK), all of which have in common that they are non-pigmented lesions in contrast to melanoma. In Sweden, the number of carcinoma diagnoses is estimated to grow in the next 20 years. Teledermatology adoption soared during the COVID-19 pandemic, increasing the need for the development of digital solutions to meet the demand. Machine learning algorithms can automate

1

Estimated number of new cases from 2020 to 2040, Males & Females, age [0-85+]
Sweden
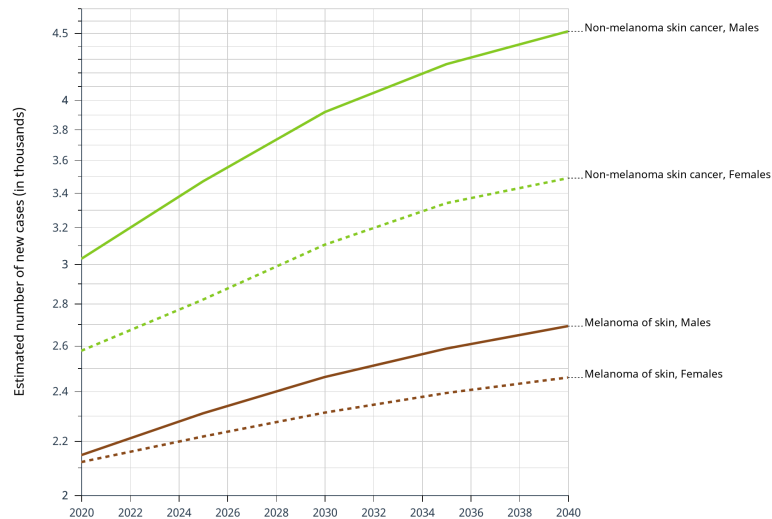
FIGURE 1.1: Estimated number of new cases from 2020 to 2040, Males & Females, age [0-85+] in Sweden. Graph obtained from the WHO cancer of tomorrow estimator [1].

the identification of skin malignancies using digital image analysis, and their diagnostic accuracy has been found to be comparable to, if not better than, dermatologists in controlled experimental settings in the last five years. In particular, clinicians with the least experience are those who benefit the most from AI-based support [7].

Even with highly experienced dermatologists and physicians, the average reported sensitivity to the diagnosis of melanoma is generally less than 80% [7]. In addition, highly qualified dermatologists are not readily available everywhere. As a result, in public health, automatic skin lesion classification at the same level of accuracy as dermatologists, if not greater, is critical.

In this project, we intend to give a different approach to the traditional one, in which our main focus will be on non-pigmented lesions rather than melanoma or pigmented lesions. The problem will be approached with simple models, useful but with obvious limitations. A machine learning model has been created to provide feedback to the dermatologist on the diagnosis of the lesion. Thus, the technology of deep learning models is brought closer to the dermatologist, leading to direct use of them in their clinics.

1.3 Objectives

The objectives of this project can be broken down as follows:

- Research of the various data sets available for the classification of skin cancer.

- Design a neural network to solve the problem of identifying non-pigmented skin lesions with dermatoscopic images.

- Analyse and validate experimentally the different results of deep learning models to solve the problem of classification of non-pigmented skin lesions.

- Identify how training a network with samples from different skin types than Sweden affects the validity of training for the Swedish skin type.

- Recognising the key elements in the image sample that lead to diagnosis.

1.4 Structure

The document is structured as follows:

- **Chapter 1:** sets the context of where we are in this project with the motivation to carry out it, some basic concepts of neural networks, and the project objectives.

- **Chapter 2:** describes the current knowledge of the matter studied through the analysis of similar or related published work, starting from a more abstract and broad topic and ending in the specific one related to the project.

- **Chapter 3:** sets a medical background for the different concepts that will be mentioned throughout the project.

- **Chapter 4:** sets a machine learning background with brief definitions of the different concepts that will be used to explain the experiments in Chapter 6.

- **Chapter 5:** exposes the different resources and tools used for the development of the project.

- **Chapter 6:** presents the steps followed in the development of deep learning models and the results obtained.

- **Chapter 7:** finally, the conclusions of the project and its future work are presented.

Chapter 2

State of the art

2.1 Image classification based on deep learning

Classification is the methodical grouping of things into groups and categories based on their characteristics. Image categorisation is accomplished by categorising the image into the appropriate category depending on the vision's content [8].

The ability to develop a computer system that resembles the human brain leads to research on artificial neural networks and deep learning [9]. In the field of machine learning, image classification plays an important role, and image processing is also widely used for image recognition and segmentation [10]. As a result, improving the classification methodology to obtain better classification accuracy is a very significant and challenging research topic. Image classification is used to close the gap between computer vision and human vision, by using images to teach the computer to recognise data in the same manner as humans do [8]. When using machine learning to solve an image classification problem, the system is made up of two parts, a feature extraction module that extracts significant features such as edges and textures, and a classification module that classifies the data based on the extracted features. However, due to some of the limitations of traditional machine learning, such as only extracting certain features and the need of having human interaction for a previous feature extraction. The deep learning methodology is introduced, in which the algorithms learn to extract high-level features from the images in an incremental manner, thanks to their architecture of hidden layers. This eliminates the need for human interaction to extract features prior to training.

In image processing as well as natural language processing, one of the most frequently used methods of deep learning is Convolution Neural Networks (CNNs) also called ConvNet. It can learn highly abstracted aspects of objects and can identify them more

efficiently than other networks with fully connected layers (FC) because of the new addition of convolutional layers with shared weights. The initial layers learn and extract the high-level features (with lower abstraction), and as the architecture grows towards the end, the deeper layers extract the low-level features (with higher abstraction). Both the classification layer and the extraction layers learn at the same time and together, making it possible to optimise both parts simultaneously and producing the output of the model more dependent on the extracted features.

Some of the major CNN architectures designed for image classification and their results in the ImageNet data set [11] are:

1. **ResNet**: [12] its architecture is different for using residual mapping instead of direct mapping [13]. The architecture is evaluated on the ImageNet data set with a depth of 152 layers. ResNet was the winner of ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC-2015) with its great performance that could reduce the top-5 error rate to 3.6% and the top-5 accuracy to 93.1%.

2. **DenseNet**: [14] the architecture extends the idea of ResNet of residual mapping, using propagation to fully connect the layers forward and backward during training [13]. The main advantages are alleviation of the vanishing gradient problem, strengthening feature propagation, encouraging feature reuse, and substantially reducing the number of parameters. DenseNet was the winner of the ILSVRC-2016 with a top-5 accuracy of 93.6% obtaining a performance similar to ResNet.

3. **Inception**: [15] the architecture demonstrates that it is possible to train high-quality networks on small data sets, thanks to a combination of lower parameter count and further regularisation with batch-normalised auxiliary classifiers and label-smoothing. The architecture has been improved multiple times, finally achieving an accuracy of 93.7% and 5.6% error top-5 in InceptionV3. A final version called Inception-Resnet (Inceptionv4) [16], in which the residual connections speed up the training process and outperforms the previous InceptionV3 by a thin margin, obtaining an accuracy of 95.3% for the top-5.

Some other architectures are **LeNet-5** from 1998, being one of the earliest CNNs available [17], **AlexNet** [18] inspired on LeNet, by Krizhevky et al. in 2012, **VGGNet** [19] is one of the most popular CNNs, introduced by Simonyan and Zisserman in 2014.

2.2 Image based diagnosis in medicine

Medical imaging is used in a variety of clinical applications, including techniques for early detection, monitoring, diagnosis, and therapy evaluation of a variety of medical problems. Therefore, there has been an emerging research field for automating these techniques with machine learning to distinguish between the presence and absence of a disease, as well as to cover the shortage of medical professionals [20]. Radiography, endoscopy, computed tomography (CT), mammography (MG), ultrasound images, magnetic resonance imaging (MRI), positron emission tomography (PET) and pathological diagnostics are only a few of the most well-known services that use imaging in medicine. The use of machine learning models in medical images can be divided into three categories, classification, segmentation, and detection. In classification, models are used to classify images into different classes, some of the most typical use cases are the identification of skin diseases in dermatology [21], the recognition of eye diseases in ophthalmology [22], pathological images such as brain cancer [23] and breast cancer [24]. In segmentation, the models attempt to contour the region of interest as an organ or anatomical structure in the image for processing. In detection, the model generally tries to identify and locate a tumour or organs in the image.

Even if the applications of machine learning to medical images are achieving promising results, there are many challenges that researchers are facing and that are slowing the progress [20]. The most relevant problems are:

1. Data quality is often lacking and it is inconsistent, such as resolution, contrast, and the existence of noise on the images.

2. Data acquisition, there are many procedures to obtain the same data and they are non-standardised, depending on the clinic.

3. Data validity, there is a need for comprehensive and valid medical notations on the samples.

4. Data sharing is complicated and has to undergo ethical approvals before being used.

Another problem is the data scarcity, which comes mainly from the sum of all the problems mentioned above, creating a bottleneck in the field. However, as said, thousands of labelled data points are not available in the medical field [20]. The technique known as transfer learning helps to solve this problem. Fixed feature extractors and fine-tuning a pre-trained network are two prominent and commonly used transfer learning methodologies. Finally, the class imbalance is another problem in which the more rare diseases

have a lower clinical incidence. Different studies have tried to solve this problem with binary classification or with the use of a previous one-class modelling [25]. The method focusses on learning models from samples belonging to a single class.

When talking about healthcare, the expected performance and accuracy of the results achieved need to meet the standards since human lives are at stake. Due to the relevance of this topic, a Healthcare workshop was held with the National Cancer Institute (NCI) and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) on 12 July 2019 [26]. The purpose of the meeting was to identify current issues and concerns related to the incorporation of machine learning into healthcare, such as trustworthiness, explainability, usability, transparency, and fairness. One of the main outcomes of the meeting was that, in order to implement machine learning models in clinical settings, the need for reliable data and methods is required but not sufficient [27]. Medical personnel need to learn to understand when to rely on the results the system outputs and when to reject them, by providing transparency of which population data the system was trained with. As well as updating the systems when new data becomes available or if the algorithm needs to be improved, necessitating of continuous monitoring.

2.3 Classification of skin images for benign and malignant cutaneous tumours using deep learning algorithms

Convolutional neural networks have achieved expert-level accuracy, sometimes even better results than experts in the case of pigmented lesions. However, in this project, we are studying the non-pigmented lesions, and these ones are harder to diagnose. The section will be divided into two parts, making a differentiation between pigmented and non-pigmented.

It is worth mentioning that different types of skin play an important role in the performance of a machine learning algorithm [28]. Algorithms used within a skin type underperform in populations independent of the training one. This was first seen in the case of the ASAM data set in which the images come mostly from East Asian skin types, and even if the algorithm was trained and validated with a good performance, it underperformed on skin lesions from white patients from the USA.

2.3.1 Pigmented

Pigmented lesions are those with melanin (which may be black, brown, grey or blue, depending on the depth in the skin) [29]. Although melanoma is not the most common

pigmented skin cancer, it is malignant and a deadly cancer. Melanoma is typically detected by a visual examination that focusses on the cutaneous lesion's area. However, there is a lot of overlap between melanomas and other skin lesions, such as nevus, making cancer categorisation and diagnosis more complex [30].

In studies [31] to detect pigmented lesions with machine learning, different types of learning techniques have been used, Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN). Most of the studies conducted focus on classifying between benign and melanoma rather than a set of different classes. Focussing our work on the results obtained using CNN, in which the deep learning models have accomplished remarkable performance in classification and segmentation. Kalouche [32] proposed a pre-trained CNN architecture that used CNN VGG-16 as the base and trained on the ISIC data set, the model obtained a 78% accuracy for melanoma lesions. Ali [33] using the LightNet architecture [34] to perform a binary classification between malignant or benign obtained on ISIC 2016 with an accuracy of 81,6%. Hasan [35] trained their own CNN and used approximately 23907 images from the ISIC archive, obtaining a total accuracy of 89.5%. Lequan [36] proposes different methods of CNN models to recognise melanoma from other benign lesions. It proposes first to segment the skin lesions to use the lesion region rather than the whole dermatoscopic image, and then classifying them between melanoma and non-melanoma. The final accuracy obtained using this method is 94.9%.

The performance of the different literature presented is varied, and many factors affect the validity of the results, such as the image types, the volume of images used, number of classes for the different types of diagnosis, and techniques utilised.

2.3.2 Non-pigmented

Non-pigmented means lesions without melanin pigmentation [29]. A non-pigmented lesion may still have pigment due to keratin (yellow or orange) or hemoglobin (red, purple, blue, or black). Both melanin and hemoglobin can produce blue and black pigmentation, which can occasionally cause confusion in the diagnosis. These lesions can lead to a broad class of tumours, including BCC and SCC, and as a group, these are the most frequent cancers that occur in light-skinned humans [37].

Similar to the pigmented studies, different types of techniques have been used. Although the number of studies carried out for only non-pigmented classes is low and they are normally included in models that classify both pigmented and non-pigmented skin lesions. For example, the case of Tschandl [21] in which they trained a CNN with multiple pigmented and non-pigmented classes and compared the results obtained from the CNN

with human experts. The accuracy of the predictions among the human raters depended on the years of experience they had, obtaining with the CNN better results in the case of beginners, but slightly worse in the case of experts. Despite the good results, the model obtained was not suitable for clinical application, as the metrics to define the accuracy do not define the accuracy in all medical settings. For example, misclassifying a life-threatening disease, for another one that left untreated, is not dangerous. They confirmed that using dermatoscopic images was better for CNN and human raters compared to clinical close-ups alone. On the other hand, Ulzii-Orshikh [38], classifies between the 3 most common non-pigmented classes (BCC, SCC and AK) and melanoma using a combination of Error-Correcting Output Codes (ECOC) and Support Vector Macine (SVM) with an existing pre-trained AlexNet leaded to a mean accuracy of 94%.

The different performances of the models of non-pigmented compared with the previous section of pigmented models can be explained knowing that the features of non-pigmented lesions are less specific than pigmented lesions, which is reflected both in the models and in the human expert raters' low accuracy.

Chapter 3

Medical background

3.1 Fitzpatrick skin phototype

The Fitzpatrick skin type (or phototype) is a scale used to classify the skin according to its reaction to exposure to sunlight [39]. The different categories depend on the amount of melanin pigment in the skin. This is determined by constitutional colour (white, brown, or black skin) and the effect of exposure to ultraviolet radiation (tanning). The Table 3.1 summarises the characteristics of each of the phototypes with its features and tanning abilities, and Figure 3.1 summarises its most characteristic colour. Skin colour is relevant to our study, since the lesions are not equally presented in the different types of skin, as non-pigmented skin carcinomas are more present in lighter skin.

TABLE 3.1: Fitzpatrick skin type

Skin Type	Features	Tanning ability
I	Pale skin	Always burns, does not tan
II	Fair skin	Burns easily, tans with difficulty
III	Darker white skin	Sometimes mild burn, tan about average
IV	Light-brown skin	Rarely burns, tans easily.
V	Brown skin	Never burns, tans very easily.
VI	Black skin	Never burns, tans very easily.

FIGURE 3.1: Illustration of the different skin colors following the Fitzpatrick skin phototype.

3.2 Dermatoscopy

Dermatoscopy or dermoscopy is a non-invasive technique for diagnosing skin lesions that refers to the examination of the skin using skin surface microscopy and is also called 'epiluminoscopy' and 'epiluminescent microscopy' [40]. Dermatoscopy requires a high-quality magnifying lens and a powerful lighting system (a dermatoscope).

A dermatoscope (also known as a dermoscope) is a handheld instrument, equipped with a magnification lens and a light source [41]. It enables viewing the subsurface morphology of cutaneous lesions, down to the depth of the superficial dermis. It enhances the diagnostic accuracy and confidence of experienced users by revealing colours and structures that are normally not visible to the unaided eye, for both pigmented and non-pigmented skin lesions. There are 2 types of dermatoscope available, Polarized (PD) and Non-polarized (NPD):

1. **Non-polarized dermatoscope (NPD)**: have a magnification lens and light-emitting diodes to provide illumination. They require direct contact of the glass plate with the skin surface and the presence of a liquid interface with a refractive index equal to or closely matching that of the skin. NPD allows visualisation of subsurface structures located in the epidermis and the dermal-epidermal junction (DEJ), but features deeper than the DEJ can hardly be visualised.

2. **Polarized dermatoscope (PD)**: also contain light-emitting diodes and a magnification lens to provide illumination like non-polarized dermatoscope. To produce cross-polarization, PDs, on the other hand, utilise two polarized filters. As a result, they do not need to come into direct contact with the skin or employ immersion liquids since it is not required to reduce the reflection of the corneal layer of the skin. PD enables visualisation of subsurface structures located at the dermal-epidermal junction (DEJ) or superficial dermis, and they are almost "blind" to the skin's surface and to structures in the superficial epidermis.

We can point out some differences between the two methods. The depth of seen structures is the fundamental difference between non-polarized dermatoscopy and polarized dermatoscopy. Although NPD is better for evaluating structures in the shallow skin layers (e.g., the superficial epidermis down to the dermo-epidermal junction), PD is better for evaluating structures in the deeper skin layers (e.g., DEJ and superficial dermis). Blood vessels (especially those located in the dermis) and vascular blush are more evident under PD since it does not require direct skin contact (due to lack of pressure effect). When compared to NPD, PD will show slightly darker brown and blue colours,

as well as greater variability in pigmentation in pigmented lesions with melanin at the DEJ or superficial dermis.

The "naked eye" evaluation alone is not enough to ensure the accurate diagnosis of skin cancer, but is important in the assessment of the gross morphological characteristics of the lesion, such as size, shape, colours, contours, and surface topography. Therefore, experts rely on dermatoscopic images to make a diagnosis.

3.3 Skin lesions and carcinomas

We can classify the different lesions between benign and malignant (carcinoma) and pigmented and not pigmented. In the following sections, the most representative types of skin lesions present in the world are explained in terms of their dermatoscopical characteristics to help us to learn how to differentiate them primarily.

3.3.1 Pigmented

The brown pigment, known as melanin, is commonly involved in pigmented lesions, causing small or large patches of the skin to appear brown, black, or even blue in colour.

3.3.1.1 Benign

Seborrheic keratoses: is a benign lesion that can be present anywhere on the skin, except mucous areas, palms and soles [42]. It normally starts to appear with ageing, especially after the 50s. The appearance of the lesion changes as it evolves. It begins as a light to dark brown oval, but as the lesion grows, it takes on the appearance of a plaque with a waxy or stuck-on aspect. They are easy to diagnose except when they are deeply pigmented and can be misjudged as melanoma and a further test must be performed. Image examples from the above can be seen in Figure 3.2.

Melanocytic nevus: is a common benign lesion that usually appears in young children and young adults and it is caused by sun exposure particularly during childhood [43]. They tend to be symmetrical and a with uniform structure, there are different types, they can be flat, flat but with a central elevation or entirely elevated. When observed under a dermatoscope they can have a cobblestone pattern, aggregated globules, network pattern or strutureless amorphus pigmentation areas [44]. Examples of the above can be found in Figure 3.3.
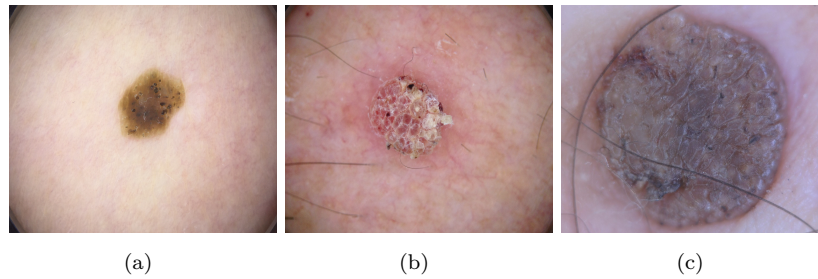
FIGURE 3.2: (a) Seborrheic keratoses in the beginning with a brown oval (b) Seborrheic keratoses evolved into the apperance of a plaque with light color (c) Seborrheic keratoses evolved into the apperance of a plaque with dark color. Images were obtained from the ISIC Archive [2].
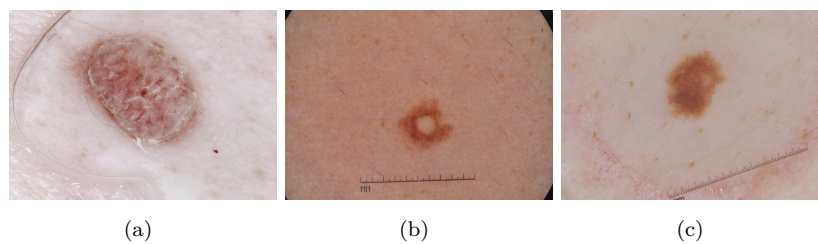
FIGURE 3.3: (a) Melanocytic nevus entirely elevated with a vascular pattern (b) Melanocytic nevus flat with a central elevation (c) Melanocytic nevus flat with an amorphous pattern. Images were obtained from the ISIC Archive [2].

Dermatofibroma: is a benign lesion with no evidence of what is the cause of them, it is thought to come from minor trauma, such as insect bites that are occasionally blamed for their appearance. They are normally located in the legs, but they can also appear in the arms and torso and are a bit depressed, like a pinch [45]. They have a different appearance depending on the type of skin, for fair skin types, it can range from pink to light brown, whereas in dark skin it can range from dark brown to black, with some appearing paler in the centre. Examples from the above are in Figure 3.4.
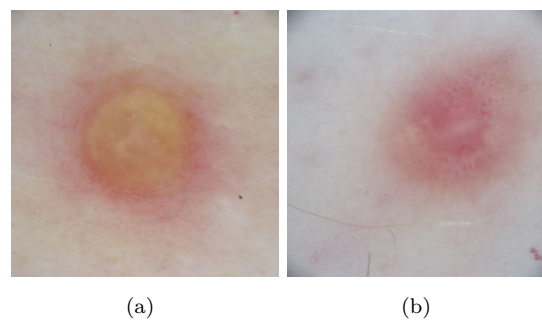
FIGURE 3.4: (a) Dermatofibroma with a pale-yellowish center (b) Dermatofibroma where the pinch can be observed and with a pink color. Images were obtained from the ISIC Archive [2].

3.3.1.2 Malignant

Melanoma: is a cancerous pigmented lesion that is less common than other types of skin carcinomas but widely known since it easily grows and spreads and for its pigmented appearance [46]. It may arise from previous benign lesions, such as acquired naevi beginning in cells known as melanocytes. It is more common for fair skin types. The dermatoscopic characteristics are extremely diverse. The early stages of melanoma are well organised, since they fall into the category from which the lesion is derived, but as time passes, it gets more disorganised, which can be seen dermatoscopically. Melanomas have dermatoscopical characteristics that indicate melanocytic origin, such as networks, brown, or black globules clumped together. In terms of shape and structure, superficial melanoma is asymmetrical and uneven and often diagnosed when they have <6mm of diameter. Examples from the above are shown in Figure 3.5.

(a)   (b)   (c)
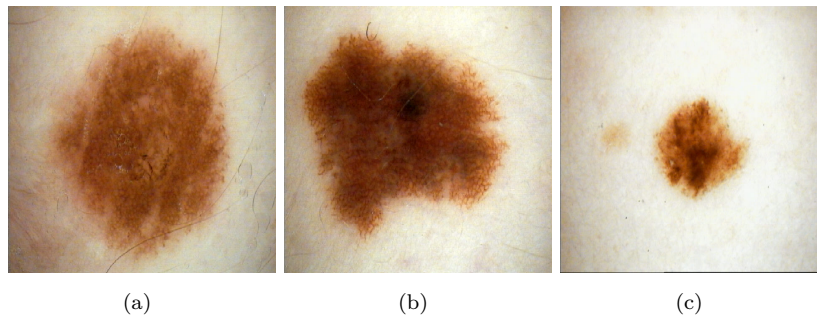
FIGURE 3.5: (a), (b) and (c) Melanomas with an irregular pigmented structure. Images were obtained from the ISIC Archive [2].

3.3.2 Non-pigmented

Non-pigmented lesions are known for the absence of melanin, giving them light colors such as red, pink, white, or skin-coloured.

3.3.2.1 Benign

Actinic keratosis (AK): is a benign lesion that commonly arises on sun-damaged skin [47]. Although it is normally not serious, it can develop into a Squamous Cell Carcinoma. The lesion appears in varying shapes depending on the part of the body it arises. It has been reported to manifest with different patterns on the face and body. The lesions found on the face dermatoscospically have what is described as the 'strawberry pattern'. Characterized for having a pink structureless background with small white areas in between that correspond to follicular openings of the skin. On the other hand,

AK non-facial does not have a specific pattern being irregular and with a surface scale typically white or yellow. Additionally, they present a rosette sign [3], that it can only be seen under polarized light and consists of 4 white dots in a clover shape and always oriented in the same angle. Examples of the above are shown in Figure 3.6

<div align="center">(a)   (b)   (c)</div>
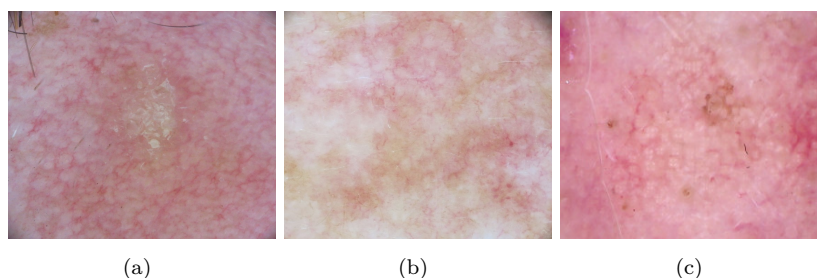
FIGURE 3.6: (a) AK facial with 'strawberry pattern'. Image from ISIC Archive [2] (b) AK non-facial. Image from ISIC Archive [2] (c) AK rosettes can be observed. Image from DermNet NZ [3].

3.3.2.2 Malignant

Basal cell carcinoma (BCC): is the most common skin cancer in the world. There are multiple types of pathologies associated with BCC and they can be either pigmented or non-pigmented. In this project, we are going to focus on the non-pigmented. BCCs frequently show pigment during dermatoscopy, with up to 30% of clinically non-pigmented BCCs displaying pigment [48]. For both pigmented and non-pigmented BCCs, the use of dermatoscopy has greatly increased diagnostic accuracy and confidence in doctors. The presence of arborizing vessels, enormous blue-gray ovoid nests, multiple blue-grey globules, leaf-like areas, spoke wheel areas, and ulceration, are dermatoscopic criteria linked with BCC [49]. Examples of the above are shown in Figure 3.7.

<div align="center">(a)   (b)   (c)</div>
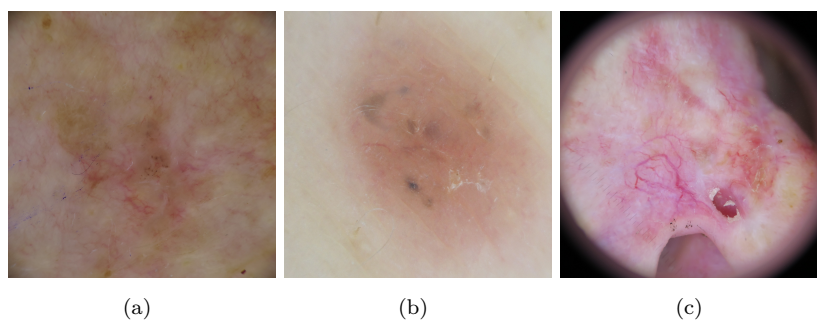
FIGURE 3.7: (a) BCC showing the presence of arborizing vessels (b) BCC with blue-gray ovoid nest present in the lesion (c) BCC showing arborizing vessels and ulceration. Images were obtained from the ISIC Archive [2].

Squamous cell carcinoma (SCC): after Basal Cell Carcinoma, it is the second most common cutaneous cancer, with a rising frequency worldwide. Actinic keratosis can

often develop into an invasive SCC. It most commonly appears on sun-exposed skin areas such as the scalp, face, neck, forearms, and dorsal hands [50]. Keratin/scales, blood spots, white circles, white structureless patches, hairpin vessels, linear irregular arteries, perivascular white halos, and ulceration are all dermatoscopic criteria for SCC [49]. Examples of the above are shown in Figure 3.8.

(a)   (b)   (c)
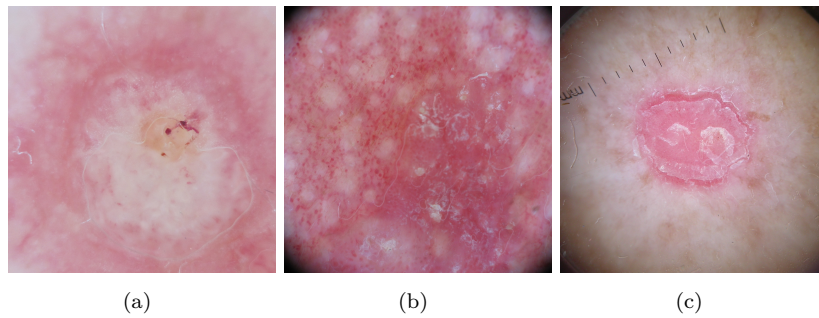
FIGURE 3.8: (a) SCC with a white structureless area and central keratin (b) SCC with white circles (c) SCC with loop vessels. Images were obtained from the ISIC Archive [2].

Chapter 4

Deep learning background

Deep learning is a subfield of Machine learning (ML) which in turn is a subset of Artificial Intelligence (AI), born from the biological neural networks of the human brain. It also present a series of characteristics similar to those of the human brain, the possibility of learning through repetition of examples, generalising the new examples to the past ones through the realisation of an abstraction of these [51]. In machine learning, there are two big categories: supervised and unsupervised learning. In supervised learning, the data used to train the model will have labelled tags, whereas in unsupervised learning the model will try to infer a behaviour in the data and look for similarities and differences in the data set. A deep learning algorithm is formed by connecting different layers. Each of the layers is composed of a certain number of neurones in which simple computations are performed, and as a whole, they form a complex algorithm.
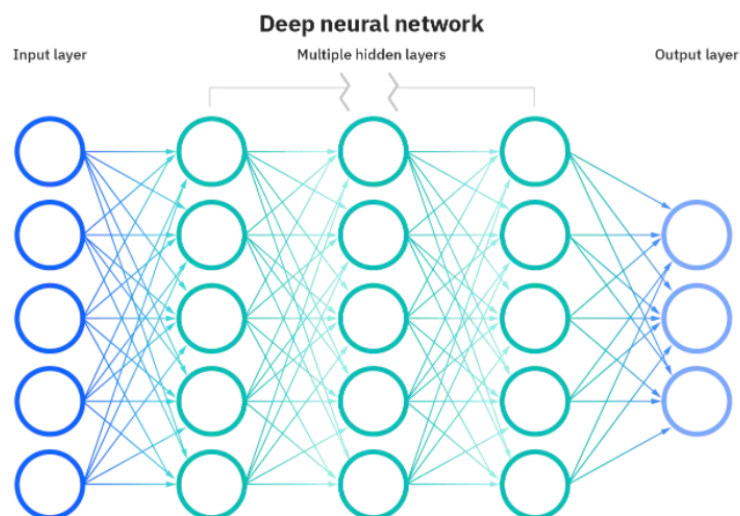
FIGURE 4.1: Neural network architecture. Image from IBM Cloud Education [4].

4.1 Deep learning concepts

In this section, we define some of the concepts used in deep machine learning models that will be mentioned throughout the document.

- **Convolutional neural network**, explained in chapter 2.1 is a neural network that uses at least one convolutional layer [52]. Used mainly for image processing, classification, segmentation.

- **Sample**, is an element of our data set. In our case, since we work with images, we refer to a single image [53].

- **Batch size**, is a hyperparameter that defines the number of samples that the network is going to work on within each iteration before updating the internal parameters of the model. The most popular sizes for this parameter are 16, 32, 64, and 128. By definition, the larger the value, the better the approximation and the training results, but it will take more time to be processed and require more memory usage. Therefore, it is recommended to use a size value as large as possible but without losing memory [53, 54].

- **Epoch**, is a hyperparameter that defines the number of times the learning algorithm will look at all the images once. Each epoch is made up of one or more batches. There is no predefined number, the value of this parameter will depend on the needs of our network [53, 54].

- **Weight**, is the coefficient for an attribute in a linear model or a connection in a deep network. The purpose of training linear models is to find the best weight for each feature. If the weight of a feature is zero, the feature does not contribute to the model [52].

- **Learning rate**, is a fundamental hyperparameter that determines how large the update of the network weights is at the end of each iteration. This determines the model's training speed and optimality; a greater rate causes the model to learn faster at the expense of achieving a set of non-optimal weights. On the other hand, a small rate will cause the model to learn more slowly but will allow it to reach a more optimal set of weights unless it reaches a local minima [55].

- **Decay**, is a hyperparameter that is responsible for adjusting the learning rate during training depending on the decay function elected. Some decay functions can be exponential, stepwise, or dependent on time [56].

- **Optimizer**, is an algorithm that modifies the parameters of the network, such as the learning rate and the weights, to reduce the losses produced while training [52]. There are different types of algorithms, each with their advantages and disadvantages.

- **Stochastic Gradient Descent (SGD) optimizer**, is an algorithm to update the parameters of the network. It is an iterative algorithm that starts from a random point and goes down the slope until it finds the lowest point of the function [52]. At each step of the training, the algorithm calculates the gradient for one observation picked at random, until the gradient converges. While the gradient has not converged the gradient will continue to be updated, the problem occurs when the function is not convex and it is possible to have reached a local minimum instead of absolute minimum [57].

- **Adam optimizer**, (Adaptive Moment Estimation) is different from the classical SGD, it combines the benefits from other two optimization methods, Adaptive Gradient Algorithm (AdaGrad) and Root Mean Square Propagation (RMSProp) [58]. Adam uses the average of the second moments of the gradients to adjust the parameter learning rates. It is a popular algorithm in machine learning since it has achieved good results in most cases.

- **Activation function**, function that defines how the weighted sum of all inputs from the previous layer is transformed into an output value for the next layer [52].

- **Sigmoid**, is an activation function, ranging between 0 and 1, normalising the output of each of the neurones, Figure 4.2, [59, 60].

$$\phi(z) = \frac{1}{1 + e^{-z}}$$
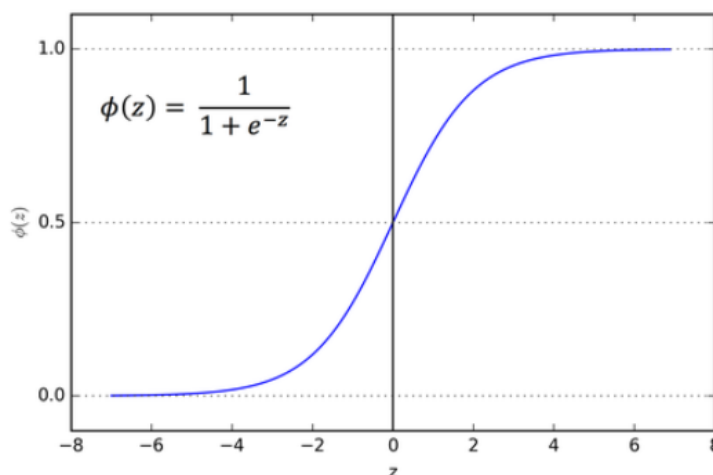
FIGURE 4.2: Plot of the sigmoid activation function, with the mathematical equation that defines it.

- **Softmax**, is an activation function, perfect for use when there is a classification between multiple classes. The function normalises the output for each class between 0 and 1 and divides it by the total sum, generating a final probability out of 1 that the input sample is from a specific class [60]. The equation 4.1 represents the function in which X_i is the k different input values.

$$f(X_i) = \frac{e^{X_i}}{\sum_{j=0}^{k} e^{X_j}} \text{ for } i = 1....k \tag{4.1}$$

- **ReLu**, Rectified Linear Unit, is an activation function, is the most common for hidden layers. In which negative values are transformed to 0.0 and if positive the value is returned for the output [52].

- **Loss function**, method to evaluate the performance of the algorithm. When training a network, we will aim the minimisation of the loss function using optimizers. In case of a deviation of the obtained results from the correct one, the resulting value of the function will increase, decreasing the success rate of the model [61].

- **Categorical cross entropy**, is a loss function, used in the training of models were there is multiple classes, using *one-hot encode* in which only one class will be active at the output. In our case, this translates to that when classifying the different non-pigmented carcinomas each image will only have one class associated to it. It is normally used together with softmax, together they output the probability of belonging to each class in total [62]. The function is defined in equation 4.2 where t_i is the corresponding target value, is the $f(X_i)$ is the i-th scalar value in the model output.

$$CE = -\sum_{i=1}^{k} t_i log(f(X_i)) \tag{4.2}$$

- **Accuracy**, is the fraction of predictions that the model got right. In our case that the classification is multiclass, the prediction is calculated as in the equation [52].

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total number of samples}} \tag{4.3}$$

- **Precision**, is the ratio of correctly predicted positive observations to the total predicted positive observations [52, 63]. If the precision is high, there will be a low number of false positives.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{4.4}$$

- **Recall**, is the ratio of correctly predicted positive observations to all observations in the actual class [52, 63].

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{4.5}$$

- **F-1 value**, is the weighted average of precision and recall. It is a very useful metric for unbalanced datasets [63].

$$\text{F1 Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4.6}$$

- **Confusion matrix**, a table that summarizes how successful a model has been in the predictions in the test set [52]. A confusion matrix has two axes: one for the predicted label and the other for the actual label. The matrix has a size of NxN, where N is the number of classes. It is a useful tool that can help find mistake patterns, for example, one class being mostly confused by another one in particular.

4.2 Pre-trained models

With transfer learning, thanks to the existence of pretrained models that have been proven to perform well in a similar task, it is possible to start training our model from a pre-existing one and not from scratch [64]. In the process, it is possible to leave frozen a number of layers, in particular the deepest ones, which are also the most specific and with a higher discriminative capacity. In addition, there is the possibility to add more layers if necessary. The choice of the model that we want to re-train to solve our need lies the compromise of the size and accuracy of the predictions. The Keras library has a wide variety of models from which we can choose. In our project, we will make use of Inception, Mobilenet, InceptionResNetV2, DenseNet for further comparison in terms of accuracies of the results [65].

Chapter 5

Resources and tools used

The project was carried out on one of the departments computers with a GPU, Nvidia Corporation GP102, Titan X. To be able to access the computer from home, an SSH connection was used.

5.1 Keras

Keras is an application programming interface (API) implemented as an open source Python library to allow the implementation of deep neural networks [66]. One of its advantages is that it is capable of running on top of Tensorflow, CNTK and Theano apart from offering support to multiple types of GPU. The API is designed to be of rapid prototyping, it is intuitive and allows high-level abstractions independent of the back-end, making it easy to understand by the user by using modular construction.

5.2 Jupyter notebook

Jupyter notebook [67] is an open source web application that allows users to create documents that contain live code, visualisations, and narrative text with the structure of runable cells. These cells can be run individually and not necessarily in sequence, this being one of its main advantages that makes it suitable for this project, since the cells can be re-runned individually, making it easy to experiment with small parts of the whole program.

5.3 Anaconda

Anaconda is an open source distribution for Python and R [68]. It was built for data scientists, and it is widely used for machine learning. Its structure of different environments allows to deploy and manage the different packages within the environment. It makes it possible to reduce the possible issues that several libraries have when working together, and also allowing the user to have different versions of a library in different environments. In this way, the project can be isolated in an environment with the preferred Python version and different versions of packages.

5.4 Dermicus

Dermicus is a CE-certified digital teledermatology platform that allows a quick and secure diagnosis of skin cancer and wounds and allows remote diagnosis [69]. Dermicus is composed of both a mobile application in which cases are created and a web platform that is used for the assessment of cases. The platform was used at Skånes University Hospital in Lund and could be accessed from both the web portal and the iOS app. The Iphone SE, was the mobile device used to collect images from patients, the phone could be coupled with a high-quality dermatoscope, the Heine iC1 [70]. When registering a new case, 4 images were obtained, 2 macroscopic and 2 dermatoscopic: the general area surrounding the lesion, a close look up to the lesion, a polarized dermatoscopic and a non-polarized dermatoscopic. Additionally, patient metadata was collected, such as age, sex, location of the lesion, time since the lesion appeared, symptoms and diagnosis. Furthermore, the web portal was used to manage the notations of the planned management of the lesion.

Chapter 6

Training and validation of the deep learning models

This chapter presents the data sets used to train the deep learning models, the methodology used, and the series of experiments carried out to test the results for the completion of the project.

6.1 Data sets

This section describes the data sets that will be used for the training and validation of the various networks. A summary of their composition is given in Table 6.1. Research into publicly available data sets was done prior to the selection of the following and can be found in the Appendix A. The following was selected based on the inclusion of the lesions in the study case, the use of dermatoscopic images, the size of the data set, and the possible similarity of skin types to Sweden.

TABLE 6.1: Number of images per class in each data set.

Data set / Class	HAM10000	BCN20000	HAM10000 & BCN20000	LU200
Actinic keratosis	327	737	1064	36
Basal Cell Carcinoma	514	3052	3566	116
Squamous Cell Carcinoma	197	431	628	22

6.1.1 HAM10000

The Human Against the Machine, HAM10000 [71], the data set is a collection of dermatoscopic images collected over a period of 20 years from different sites, the Department

of Dermatology at the Medical University of Vienna, Austria, and the skin cancer practice of Cliff Rosendahl in Queensland, Australia. The data set consists of 10015 images divided into the following categories: basal cell carcinoma, dermatofibroma, melanoma, pigmented benign keratosis, melanocytic nevus, vascular lesion, actinic keratosis, and squamous cell carcinoma in which the diagnoses are 53.3% of biopsy-proven samples. In our particular case, only the non-pigmented classes were used, BCC, SCC and AK.

The data set is publicly available in the ISIC Archive [2] and has been part of the ISIC-2018 [72] and 2019 [73] challenge. All data contained in the HAM10000 data set has received the institutional ethical approval (CC BY-NCND 4.0).

6.1.2 BCN20000

The BCN20000 data set [74] was created by *Hospital Clínic in Barcelona* from lesions captured between 2010 and 2016. The data set consists of 19424 dermatoscopic images from 5583 lesions, including lesions that are difficult to diagnose, such as those found in the nails and mucosa. The data set was divided into the following categories: nevus, melanoma, basal cell carcinoma, seborrheic keratosis, actinic keratosis, squamous cell carcinoma, dermatofibroma, vascular lesion and 'other' (lesions not contained in any other category). Furthermore, each image was paired with metadata on the lesion location, age and sex of the patient. In our particular case, only the non-pigmented classes, BCC, SCC, and AK, were used.

The data set was part of the ISIC-2019 challenge [73] and is publicly available in the ISIC Archive [2]. All data contained in the BCN20000 data set has received the necessary institutional ethics approval (HCB/2019/0413).

6.1.3 LU200

The LU200 data set was created by the Lund Univesity Hospital from lesions captured during 2 weeks in March 2022. The data set is composed by 348 dermatoscopic and macroscopic images from 87 lesions. For each lesion there are two macroscopic images, one of the general area and another with a close-up, then two dermatoscopic images polarized and non-polarized. The data set is divided into basal cell carcinoma, squamous cell carcinoma, and actinic keratosis. Each image has metadata associated with the location of the lesion, the age, sex of the patient, the time of the lesion, and the procedure that was followed.

The data is not publicly available, and the use of it was granted for research purposes to Lund University.

6.2 Experiments and results

The experiments can be divided according to the data set used and the architecture from which the model will be trained. The model architectures used are DenseNet201, Mobilenet and InceptionResNetV2 and using the technology of transfer learning, we can retrain them to solve our problem. The complete architecture for each experiment is given in the Appendix B. The experiments were based on the following configuration in which the hyperparameters have been defined as those used by the Adam optimiser [58] and a batch size of 16 or 32, we vary the learning rate in the different experiments and in addition for the last layer we will use softmax as the activation function and categorical cross-entropy as the loss function. We use Keras callbacks and as a stopping criterion (*EarlyStopping*) to avoid overfitting the model and class weight balance to compensate for the different amounts of images in each class. The images are resized to 224x224 pixels and divided between 70% for train and 15% each for test and validation. Since the total number of images was not sufficient for training, two types of image augmentations were implemented. First, we tried the built-in image augmentation from Keras, which provides real-time data augmentation ensuring that the model receives new input data in every epoch. Secondly, a previous manual augmentation in which all classes were manually assigned the same fixed number of images for training, validation, and testing. The modifications applied to the images were chosen based on the fact that the lesion was not deformed or cut out of the image. We applied rotation up to 260 degrees and both vertical and horizontal flipping and minor shifting. A complete summary of the experiments and the parameters that were characterised in their training can be found in Table 6.2. Furthermore, we will use the Zero rate, it is the accuracy that the model would obtain if classifying always to the most frequent class. We will use this rate to demostrate that the model is learning and that it is skilled in solving the problem rather than just going with the most probable class.

6.2.1 HAM10000 training

In the HAM10000 data set only using the three non-pigmented classes available, the total number of images per class were 514 BCC, 327 AK and 197 SCC. Therefore, the Zero rate classifier of this data set will be 49% if we do not use our previous augmentation in which the classes are balanced. We present the most relevant models and the parameters that characterise them. Additionally, a summary of each of the experiment results is presented in Table 6.3.

- **Experiment 1**: Transfer learning from the InceptionResNetV2 model in which all imported layers were trainable. After the imported model the following layers

TABLE 6.2: Summary of experiments and the parameters used for training.

	Architecture	Layers trainable	Image augmentation	Batch	Epochs	Early Stopping
HAM10000						
Exp. 1	InceptionRes-NetV2	All	Built-in	16	20	-
Exp. 2	DenseNet201	All	Own	32	40	-
Exp. 3	Mobilenet	All	Built-in	32	40	-
BCN20000						
Exp. 1	DenseNet201	Last 10	Own	32	30	17
Exp. 2	DenseNet201	All frozen	Own	32	40	18
Exp. 3	Mobilenet	All frozen	Own	32	40	-
HAM & BCN						
Exp. 1	DenseNet201	Last 20	Built-in	32	30	-
Exp. 2	DenseNet201	Last 10	Built-in	32	40	23
Exp. 3	DenseNet201	All frozen	Built-in	32	40	-

TABLE 6.3: HAM10000 Summary Table of results.

Experiment	Training		Validation		Testing	
	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	0.88	0.29	0.83	0.51	0.76	0.62
2	0.98	0.04	0.54	1.08	0.78	0.75
3	0.67	0.74	0.64	0.78	0.62	0.84

were added: average pooling, 3 fully connected layers with their respective dropout layer with a value of 0.5 to prevent overfitting and the flattening layer before the final prediction layer. This experiment used Keras built-in augmentation, learning rate 0.0001, batch size of 16, and a total of 20 epochs. After training, we obtained a training loss of 0.29 and an accuracy of 0.88, validation loss of 0.51 and an accuracy of 0.83, and a testing loss of 0.62 and an accuracy of 0.76. The model has improved the baseline set by the Zero rate by 27%. In Figure 6.1 on the left it can be seen that training and validation accuracy and loss are both increasing and decreasing at a similar pace. On the top right of the figure, we have the classification report in which we can see that for the SCC class the precision is low while on the other hand its recall is high, remembering the equation from chapter 4 we can say that the model is incorrectly classifying samples from other classes into the SCC class. For the AK class, we have high precision but low recall, meaning that the model does not correctly classify AK samples and incorrectly classifies them into another class. In this case, as can be seen in the confusion matrix below, they are classified as SCC.

- **Experiment 2**: Transfer learning from the DenseNet201 model in which all imported layers were trainable. After the imported model, the following layers were

FIGURE 6.1: HAM10000, experiment 1: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.

added: average pooling, 2 fully connected layers with their respective dropout layer with a value of 0.5 to prevent overfitting, and the flattening layer before the final prediction layer. This experiment used our own previous image augmentation, learning rate 0.0001, batch size of 32, and a total of 40 epochs. After training, we obtained a training loss of 0.04 and an accuracy of 0.98, a validation loss of 1.08 and an accuracy of 0.54, and a testing loss of 0.75 and an accuracy of 0.78. In Figure 6.2 on the left, we can see that the model is overfitted and has reached almost maximum accuracy during training, but when validated, the model abruptly varies its accuracy between consecutive epochs. In the confusion matrix, it can be seen that most of the classification errors are made for class AK, with a low recall score, this model confuses this class with the other 2 in 40% of the cases.

- **Experiment 3**: Transfer learning from the Mobilenet model in which all imported layers were trainable. After the imported model the following layers were added: average pooling, 3 fully connected layers with their respective dropout layer with a value of 0.5 to prevent overfitting and the flattening layer before the final prediction layer. This experiment used Keras built-in augmentation, learning rate 0.0001, batch size of 32, and a total of 40 epochs. After training, we obtained a training loss of 0.74 and an accuracy of 0.67, validation loss of 0.78 and an accuracy of 0.64, and a testing loss of 0.84 and an accuracy of 0.62. The model has improved the baseline set by the Zero rate by 13%. In Figure 6.3 on the left, the accuracy and loss graphs are very spiky for both training and validation, which could mean that the weights in that particular epoch are in a bad local minima, this could be

FIGURE 6.2: HAM10000, experiment 2: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.

solved by lowering the learning rate or changing the optimiser. The same problem defined for experiment 1 is present for this one, low precision for SCC class and low recall for AK class.

FIGURE 6.3: HAM10000, experiment 3: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.

6.2.2   BCN20000 training

In the BCN20000 data set only using the three non-pigmented classes available, the total number of images per class was 3052 BCC, 737 AK and 431 SCC. The Zero rate classifier of this model is 72% if we do not use our previous augmentation in which the classes are balanced. We present the most relevant models and the parameters that characterise them. In addition, a summary of each of the experiment results is presented in Table 6.4.

TABLE 6.4: BCN20000 Summary Table of results.

Experiment	Training		Validation		Testing	
	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	0.99	0.01	0.83	0.47	0.86	0.36
2	0.90	0.24	0.78	0.64	0.79	0.48
3	0.84	0.42	0.74	0.62	0.76	0.58

- **Experiment 1**: Transfer learning from the DenseNet201 model in which the 10 last imported layers were trainable. After the imported model the following layers were added: average pooling and the flattening layer before the final prediction layer. This experiment used our own previous image augmentation, learning rate 0.0001, batch size of 32, a total of 30 epochs with *EarlyStopping* in epoch 17. After training, we obtained a training loss of 0.01 and accuracy of 0.99, validation loss of 0.47 and accuracy of 0.83, and testing loss of 0.36 and accuracy of 0.86. In Figure 6.4 on the left it can be seen that the model is clearly overfitting, reaching an almost perfect score and plateu in training from an early epoch with a difference of 15% with validation before the model was early stopped. In the classification report we can see that for the AK class the precision is low when compared to the other two, we can say that the model is incorrectly classifying samples from other classes into the AK class, in particular the SCC as can be seen in the confusion matrix. For the SCC class, we have high precision but low recall, meaning that the model does not correctly classify SCC samples and incorrectly classifies them into another class, in particular AK.

- **Experiment 2**: Transfer learning from the DenseNet201 model in which all imported layers are frozen. After the imported model, the following layers were added: average pooling, 3 fully connected layers with their respective dropout layer with a value of 0.5 to prevent overfitting, and the flattening layer before the final prediction layer. This experiment used our own previous image augmentation, learning rate 0.0001, batch size 32, a total of 40 epochs with *EarlyStopping* in epoch 18. After training we obtained a training loss of 0.24 and accuracy of 0.90,
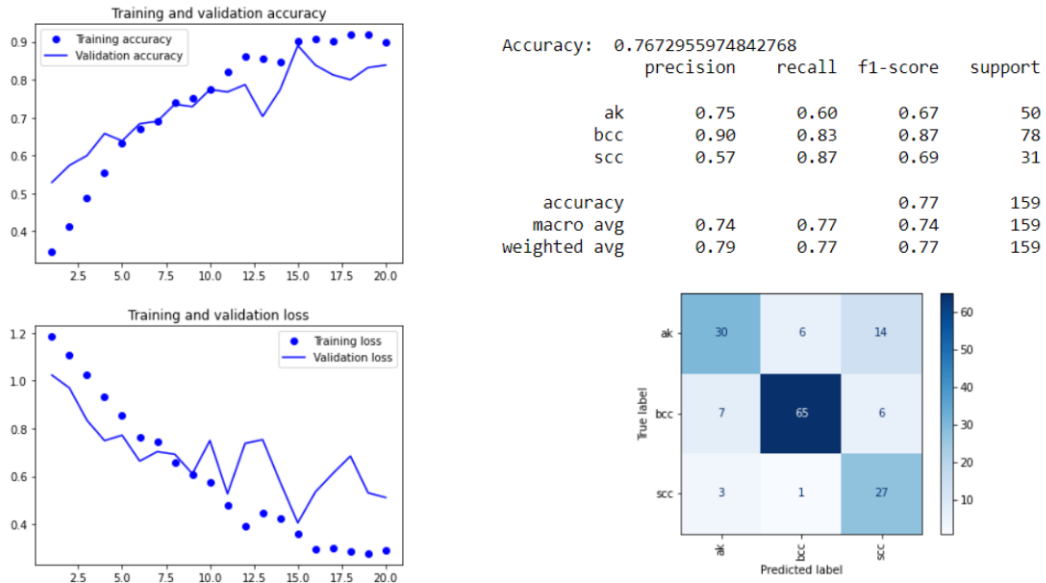
FIGURE 6.4: BCN20000, experiment 1: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.

validation loss of 0.64 and accuracy of 0.78, and testing loss of 0.48 and accuracy of 0.79. In Figure 6.5 on the left it can be seen that the model is overfitting, reaching an almost perfect score in training with a difference of 10% with validation. For the AK and SCC classes in the classification report, they have the same value for precision and recall, in which both classes are incorrectly classified as belonging to another one.

- **Experiment 3**: Transfer learning from the Mobilenet model, in which all imported layers are frozen. After the imported model the following layers were added: average pooling and the flattening layer before the final prediction layer. This experiment used our own previous image augmentation, learning rate 0.0001, batch size of 32, a total of 40 epochs. After training we obtained a training loss of 0.42 and accuracy of 0.84, validation loss of 0.62 and accuracy of 0.74, and testing loss of 0.58 and accuracy of 0.76. In Figure 6.6 on the left it can be seen that the model is clearly overfitting, reaching an almost perfect score in training with a difference of 10% with validation. In the confusion matrix, it can be seen that most of the classification errors are made for class SCC, in which this one is confused with AK in 26% of the cases.
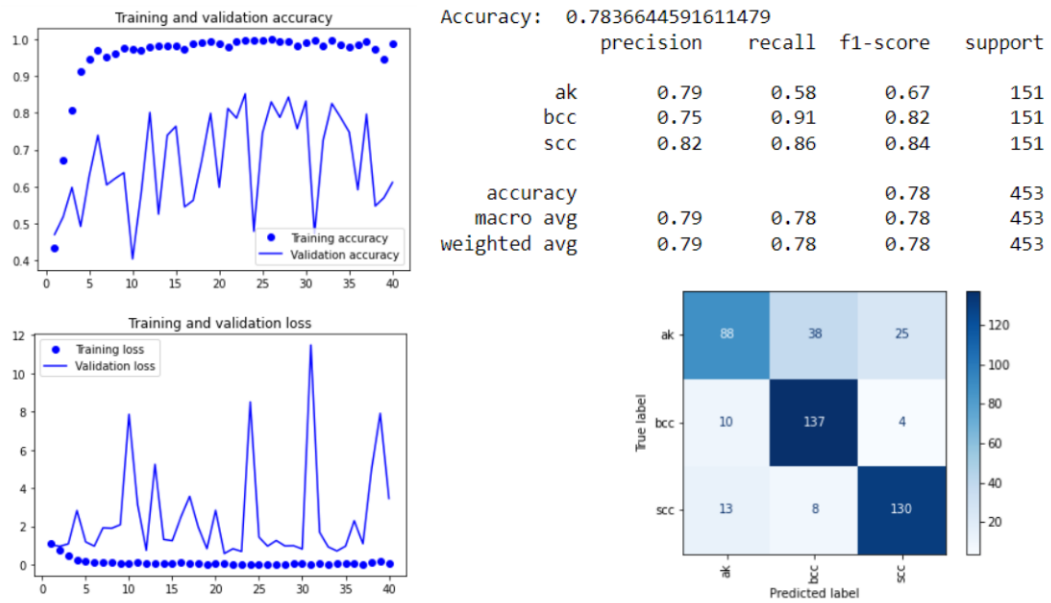
FIGURE 6.5: BCN20000, experiment 2: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.
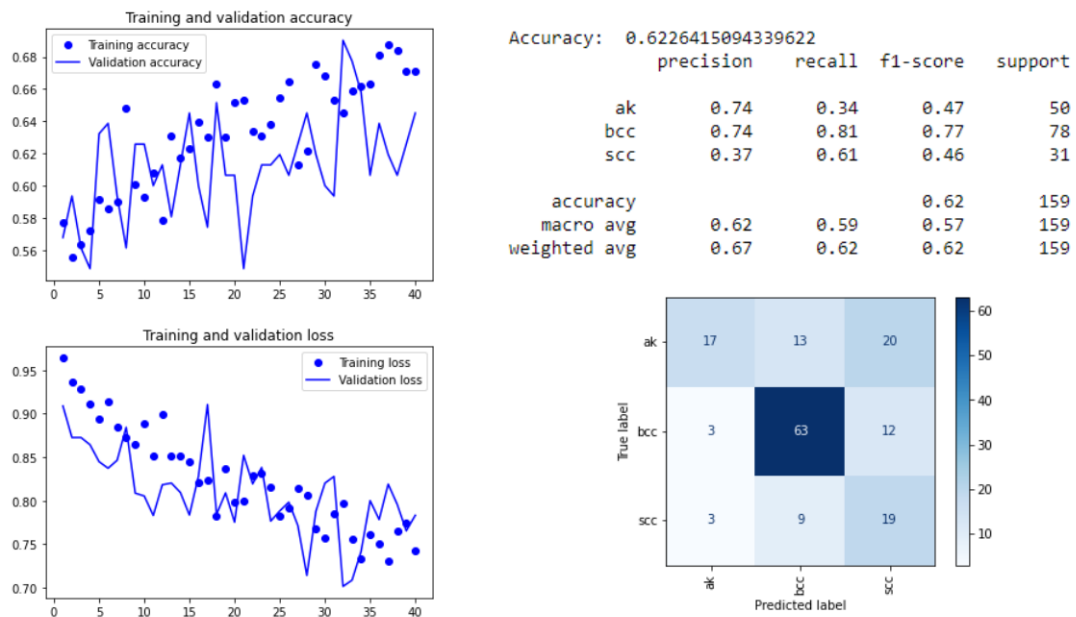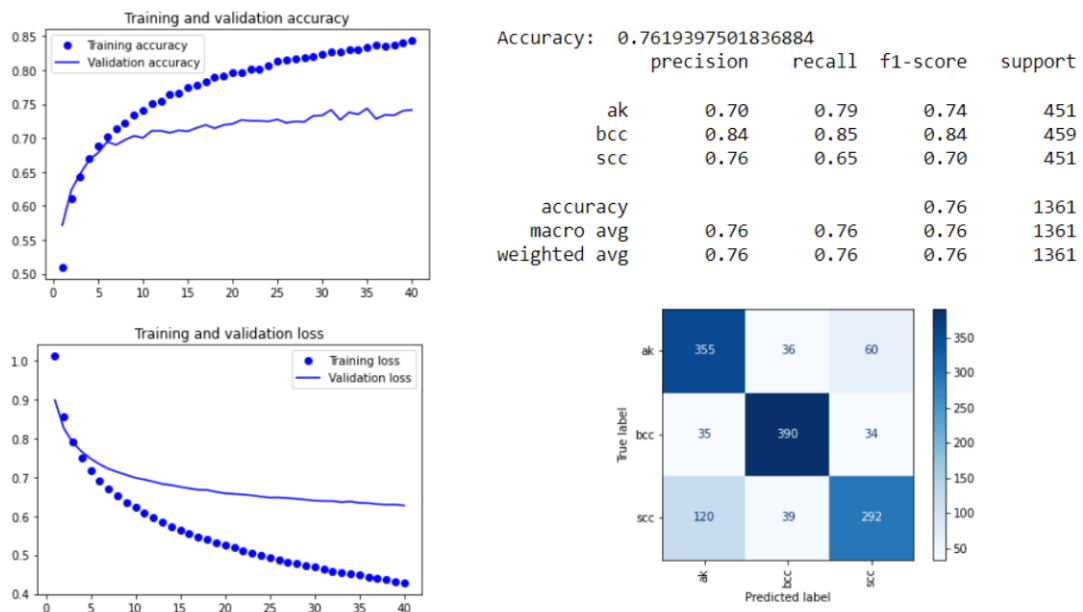

FIGURE 6.6: BCN20000, experiment 3: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.

6.2.3 HAM10000 and BCN20000 training

In this case, a new data set was created with the combination of HAM and BCN. With a total images per class of 3566 BCC, 1064 AK and 628 SCC. The Zero rate classifier of this model is 67% if we do not use our previous augmentation in which the classes are

balanced. We present the most relevant models and the parameters that characterise them. In addition, a summary of each of the experiment results is presented in Table 6.5.

TABLE 6.5: HAM10000 and BCN20000 Summary Table of results.

Experiment	Training		Validation		Testing	
	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
1	0.93	0.19	0.79	0.55	0.82	0.53
2	0.89	0.31	0.71	0.66	0.79	0.54
3	0.65	0.75	0.65	0.72	0.67	0.70

- **Experiment 1**: Transfer learning from the DenseNet201 model in which the 20 last imported layers were trainable. After the imported model the following layers were added: average pooling and the flattening layer before the final prediction layer. This experiment used Keras built-in augmentation, learning rate 0.0001, batch size of 32, and a total of 30 epochs. After training, we obtained a training loss of 0.19 and an accuracy of 0.93, a validation loss of 0.55 and an accuracy of 0.79, and a testing loss of 0.53 and an accuracy of 0.82. The model has improved the baseline set by the Zero rate by 15%. In Figure 6.7 on the left it can be seen that the model is clearly overfitting, reaching an almost perfect score in training with a difference of 14% with validation. In the classification report, we can see that for the SCC class the precision is low, we can say that the model is incorrectly classifying samples from other classes into the SCC class, in particular the AK as can be seen in the confusion matrix. For the AK class, we have high precision but low recall, meaning that the model does not correctly classify AK samples and incorrectly classifies them into another class, in particular SCC.

- **Experiment 2**: Transfer learning from DenseNet201 model in which the 10 last imported layers were trainable. After the imported model the following layers were added: average pooling and the flattening layer before the final prediction layer. This experiment used Keras built-in augmentation, learning rate 0.0001, batch size of 32, and a total of 40 epochs with *EarlyStopping* in 23 epochs. After training, we obtained a training loss of 0.31 and accuracy of 0.89, validation loss of 0.66 and accuracy of 0.71, and testing loss of 0.53 and accuracy of 0.79. The model has improved the baseline set by the Zero rate by 12%. In Figure 6.8 on the left it can be seen that the model is clearly overfitting, reaching an almost perfect score in training with a difference of 18% with validation. Although the model continues to learn in each epoch the validation is stable with little spikes in the accuracy. In the classification report, we can see that we have the same problem of low precision and recall as described on experiment 1 in the same classes.
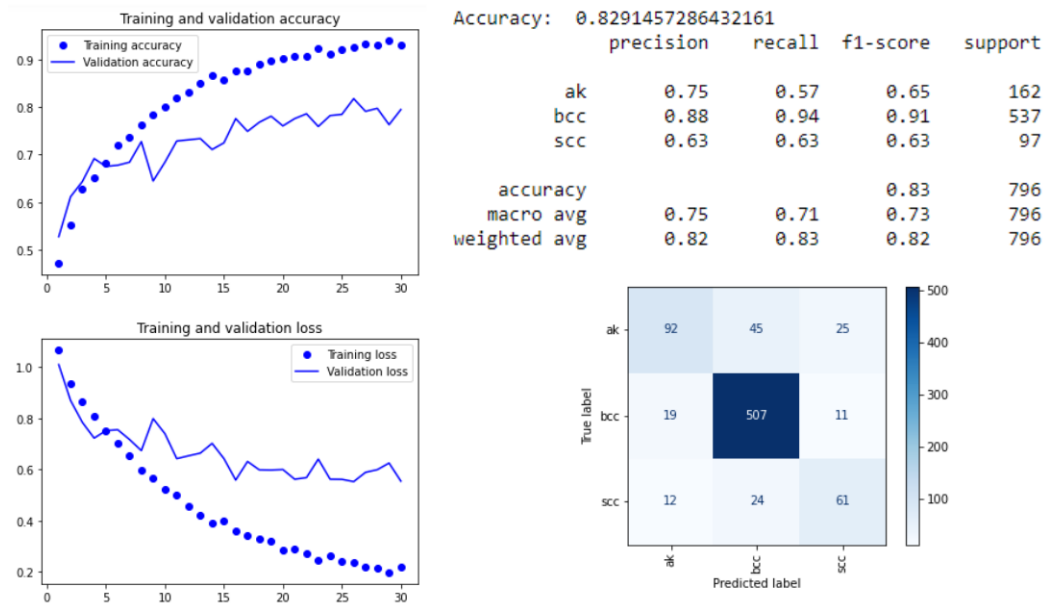
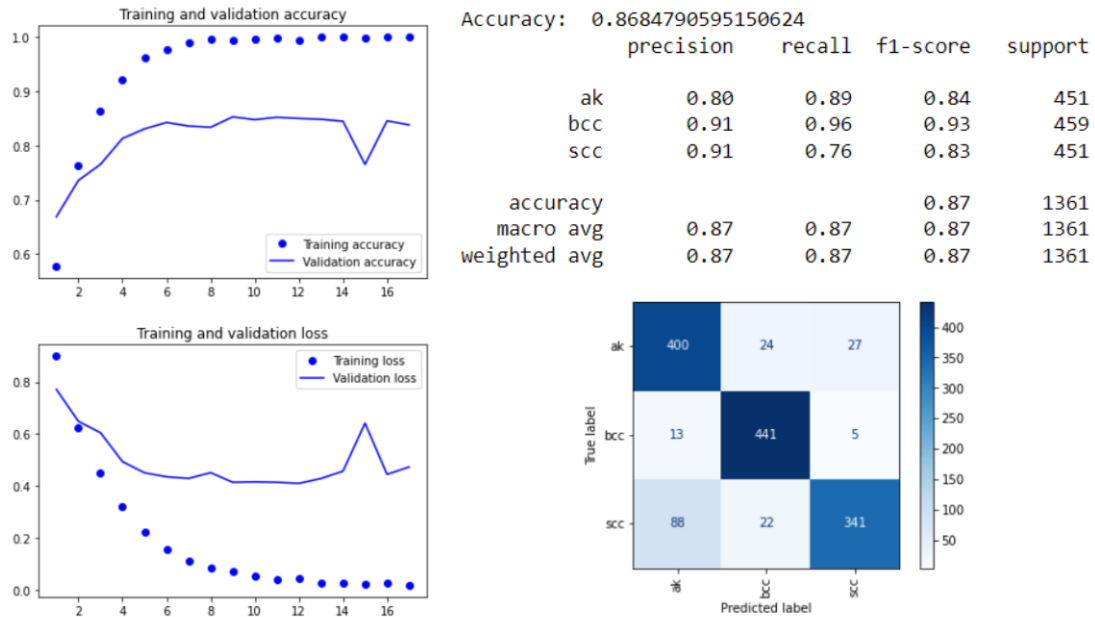FIGURE 6.7: HAM10000 and BCN20000, experiment 1: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.
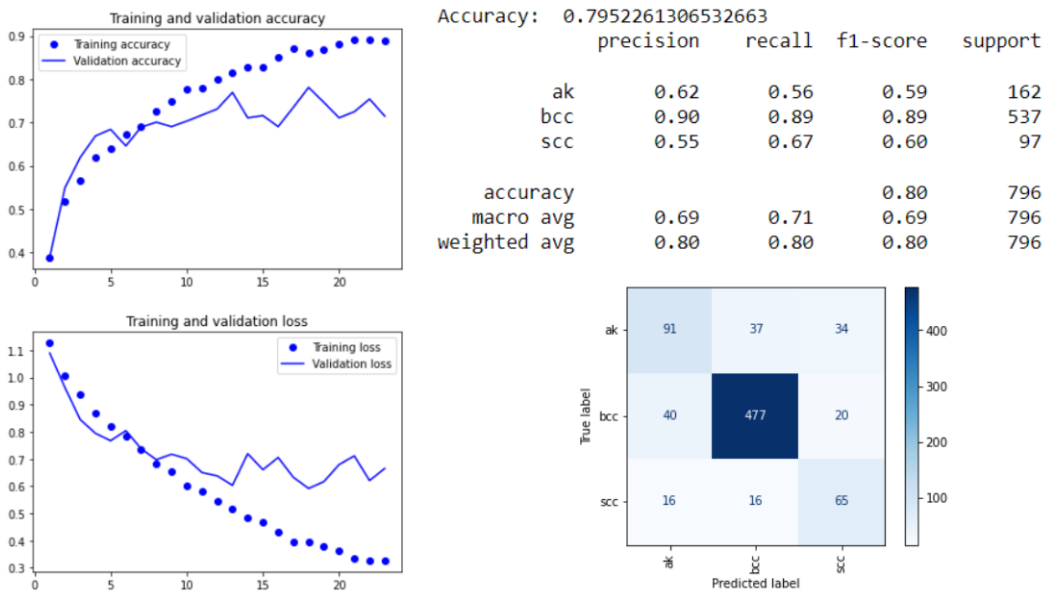
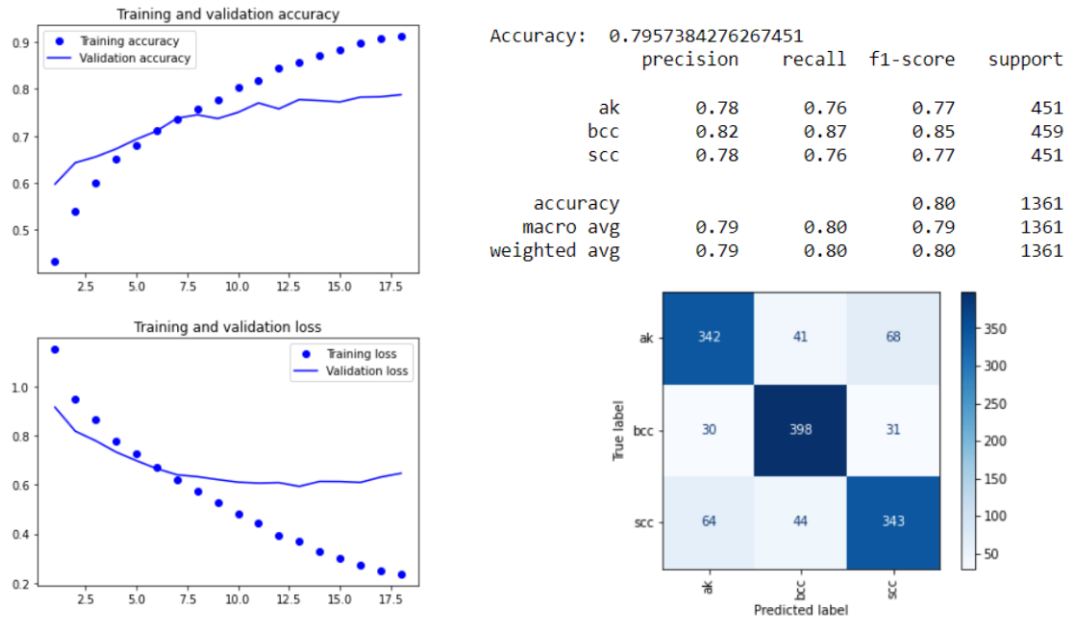FIGURE 6.8: HAM10000 and BCN20000, experiment 2: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.

- **Experiment 3**: Transfer learning from DenseNet201 model in which all imported layers were froze. After the imported model, the following layers were added: average pooling, 3 fully connected layers with their respective dropout layer with a value of 0.5 to prevent overfitting, and the flattening layer before the final prediction layer. This experiment used Keras built-in augmentation, learning rate 0.0001, batch size of 32, and a total of 40 epochs. After training, we obtained

a training loss of 0.75 and an accuracy of 0.65, a validation loss of 0.72 and an accuracy of 0.65, and a testing loss of 0.70 and an accuracy of 0.67. The model has not improved the baseline set by the Zero rate. However, as can be seen in Figure 6.9 on the left, this model has fixed the previous problem that other models had with overfitting. However, the accuracy is reduced, and the classification worsens for all classes.
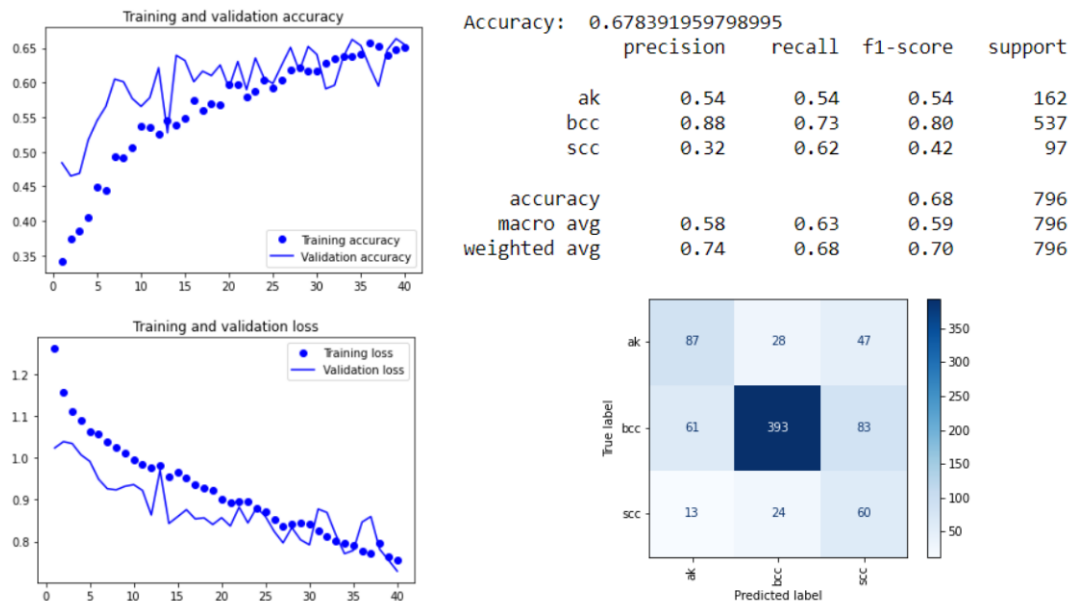
FIGURE 6.9: HAM10000 and BCN20000, experiment 3: Left: Training and validation accuracy and loss graphs per epoch. Top right: Testing classification report. Bottom right: Testing confusion matrix.

6.3 Networks versatility

The most significant trained networks were tested with the other data sets, the model chosen was selected based on the accuracy and consistency of their results. For the HAM10000 data set the model selected is from experiment 1. For the BCN20000 data set, the model from experiment 2. For the HAM10000 and BCN20000 data set, the model from experiment 1. The results of the validations are shown in Table 6.6, and Figures 6.10, 6.11 and 6.12 represent the different confusion matrices.

- **HAM10000**: when the model is tested with the other data sets, it performs poorly, obtaining low accuracies and high losses with a clear tendency to classify the samples as belonging to the SCC class as seen in Figure 6.10.

- **BCN20000**: when the model is tested with the HAM10000 data set the accuracy is 32% worse than a random classifier and classifies the samples as belonging to

BCC the most probable class, as seen in (a) of Figure 6.11. For the LU200 data set, the accuracy obtained is 40% and the predictions are mainly for the BCC and SCC classes.

- **HAM10000 BCN20000**: when the model is tested with the HAM10000 data set, the class BCC is correctly classified but for AK and SCC, the predictions are equally divided for all classes. When tested for BCN20000 the results are not far from the original model since the BCN20000 was the data set that contributed more to this model, but there is a strong preference to predict as BCC, which can also be seen for the LU200 data set, Figure 6.12.

TABLE 6.6: Summary of validation results.

Dataset / Model	HAM10000		BCN20000		LU200	
	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
HAM10000	0.78	0.76	0.34	8.45	0.18	9.08
BCN20000	0.32	2.92	0.79	0.48	0.40	1.40
HAM10000 & BCN20000	0.65	1.07	0.73	0.81	0.66	1.63

(a) (b) (c)
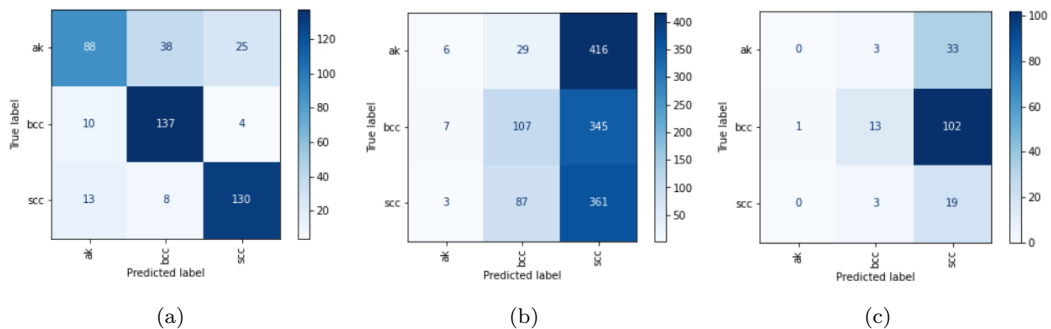
FIGURE 6.10: Confusion matrices of the testing of HAM10000 in the following data sets: (a) HAM10000 (b) BCN20000 (c) LU200.

(a) (b) (c)
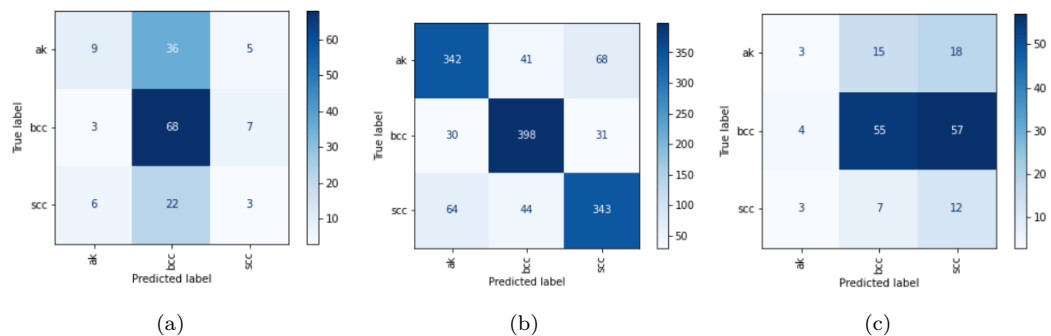
FIGURE 6.11: Confusion matrices of the testing of BCN20000 in the following data sets: (a) HAM10000 (b) BCN20000 (c) LU200.
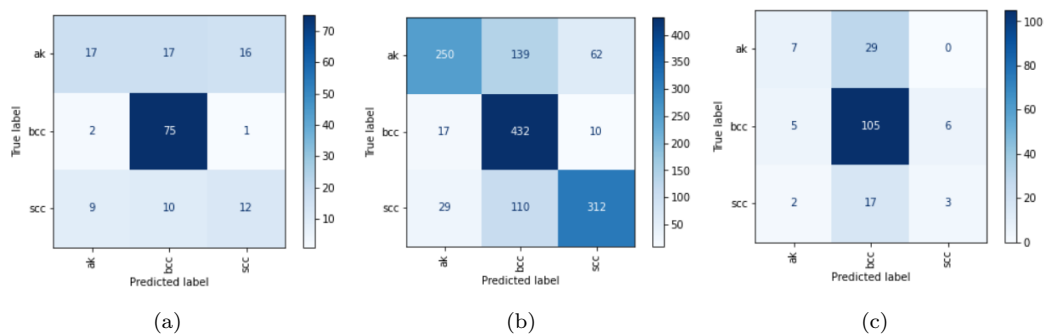
FIGURE 6.12: Confusion matrices of the testing of HAM10000 and BCN20000 in the following data sets: (a) HAM10000 (b) BCN20000 (c) LU200.

6.4 Grad-CAM

The Grad-CAM algorithm is an additional experiment carried out to determine what the neural network was focusing on to make a prediction, whether the area was correct or not. This method includes the production of heat maps that represent activation classes in the received input image. The activation class is associated with a particular output class. These classes can be used to show the importance of each pixel in relation to the class concerned by increasing or reducing the intensity of the pixel. The heatmap is created from the last convolutional layer in which the final prediction is made, and the algorithm will output the corresponding heatmap and the superimposed map on the original image [75].

The model used for Grad-CAM testing is the model from experiment 1 with the mix data set of HAM10000 and BCN20000. Grad-CAM was applied for each own testing data set and also for LU200. On the first one, some examples are presented in Figures 6.13 and 6.14 for correct classification and incorrect one respectively. From the images it can be seen that the model is not focussing directly on the lesion and more on the surrounding skin. These could be the "white clues" that are present and that dermatologists also use to base their diagnosis. The set of images in Figure 6.15 are taken from the LU200 data set and were correctly classified by the model in which the focus of the model in all images is mainly on the lesion. In contrast, in Figure 6.16 the images presented were incorrectly classified by the model, in this case the images taken with the NP dermatoscope had a glare that confused the model for (b) and (d) and focused on them instead of on the lesion. (a) and (c) are incorrect classifications from the model, which (c) can be explained as most of the crust lesions in the BCN20000 data set belonged to SCC, a crust lesion is often defined as SCC by the model. Images (b), (c) and (d) have in common that the model is looking at the black border from the image, to make a decision. These borders lack information about the type of lesion, indicating that the

model is not adapting to the new form of the data set and that possibly adding image preprocessing to eliminate the border would remove the bias.

To test the constancy of the model, we carried out a test in which parts of the images were occluded leaving only visible those parts in which the model based the decision when having the full picture. Then the Grad-CAM algorithm was run in the new occluded images to compare. In Figure 6.17, the image on (a) is the complete image used for the test in which the lesion is at the centre of the image and the model has focused on the outer borders of the lesion. When the image is cut in (b) the lesion is on the bottom left corner and the model has again focused on the border of the lesion to make the prediction, in this case again the correct one. In Figure 6.18, the image on (a) is the complete image with the lesion in the centre that belongs to the AK class, is correctly classified, and the model is focused below the white area of the lesion. However, when the lesion is cut in (b), the model concentrates directly on the white area of the lesion, leading it to miss classify it as BCC class.

6.5   Discussion

The study has shown that a CNN is capable of classifying non-pigmented lesions from dermatoscopic images. In our experimental setting, we did not use important metadata, such as age, location of the lesion, and patient history, deviating from the experiments of a real clinic diagnosis, in which dermatologists have access to these data when making a diagnosis. Our data suffer from class imbalance in which some lesions are more common, such as BCC, being over-represented. While AK and SCC had a lower number of samples for all data sets used, as seen in Table 6.1. This caused bias in the predictions, towards BCC, the problem was mitigated when applying image augmentation prior to training. The majority of the models presented from the experiments had similar problems when classifying the AK and SCC classes. This makes sense since AK is a premalignant condition that can lead if untreated to SCC. Although the data sets were one of the largest available for non pigmented dermatoscopic lesions, it is a small data set compared to others in the world of deep learning. After researching the publicly available data sets, the two data sets used for the project were HAM10000 and BCN20000. Each of them different in terms of the method of obtaining the images having clear differences in how the image was presented (e.g. black circle on the borders of the image or not, size of the circle, etc.) and the skin type surrounding the lesion. HAM10000 with more representation of Central Europe and BCN20000 of Southern Europe skin types. Therefore, when testing the versatility of the model with another data set than that
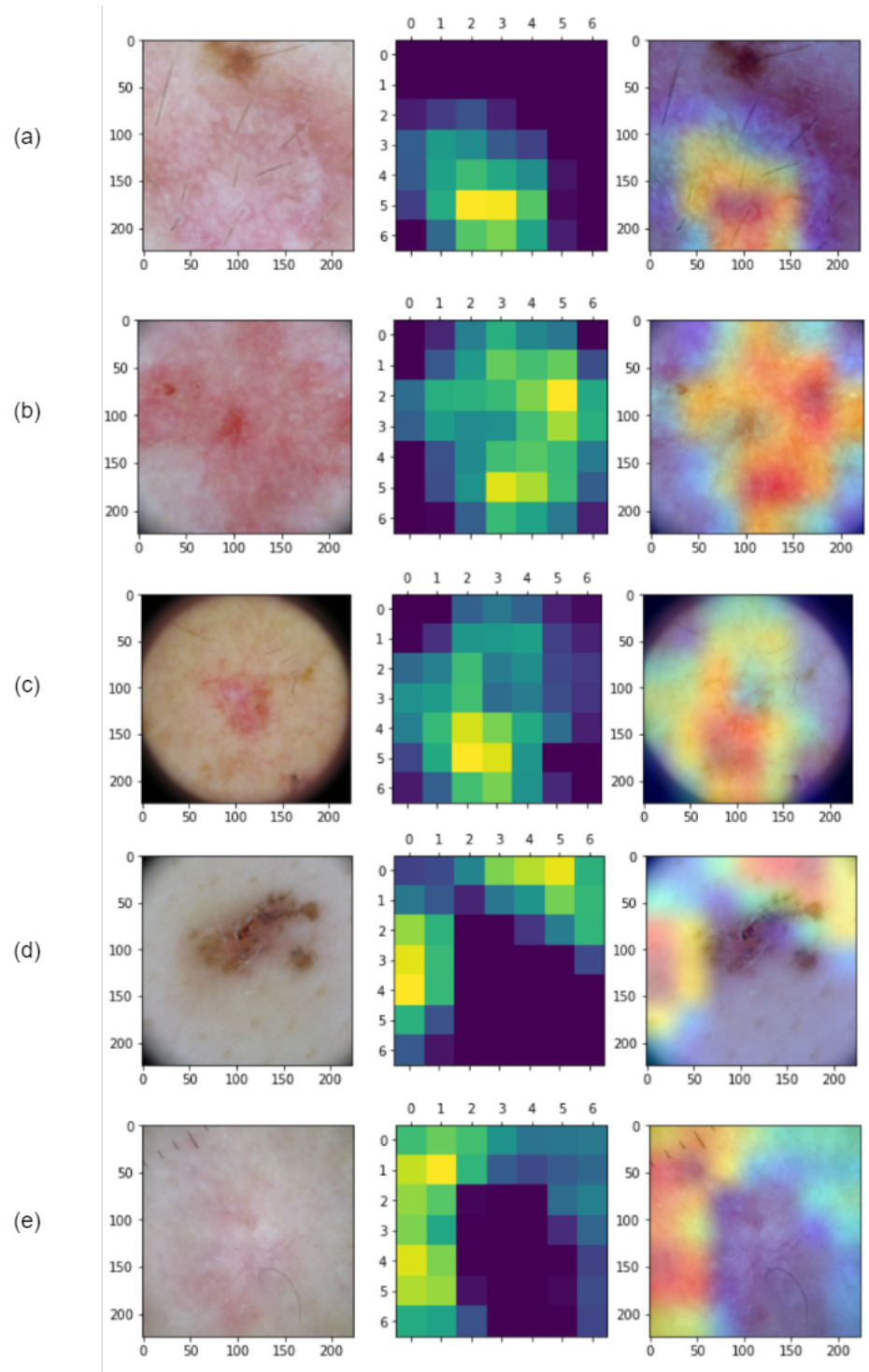
FIGURE 6.13: Visual explanations of correct classifications of the HAM10000 and BCN20000 data set via gradient based localization. First column is the original image, second column represents the heatmap and third column is the superimposed heatmap on the image. (a) AK. (b) AK. (c) BCC. (d) BCC. (e) SCC.

used for training, in Section 6.3. The models performed poorly for the reasons stated above, being unclear which one plays a larger role.
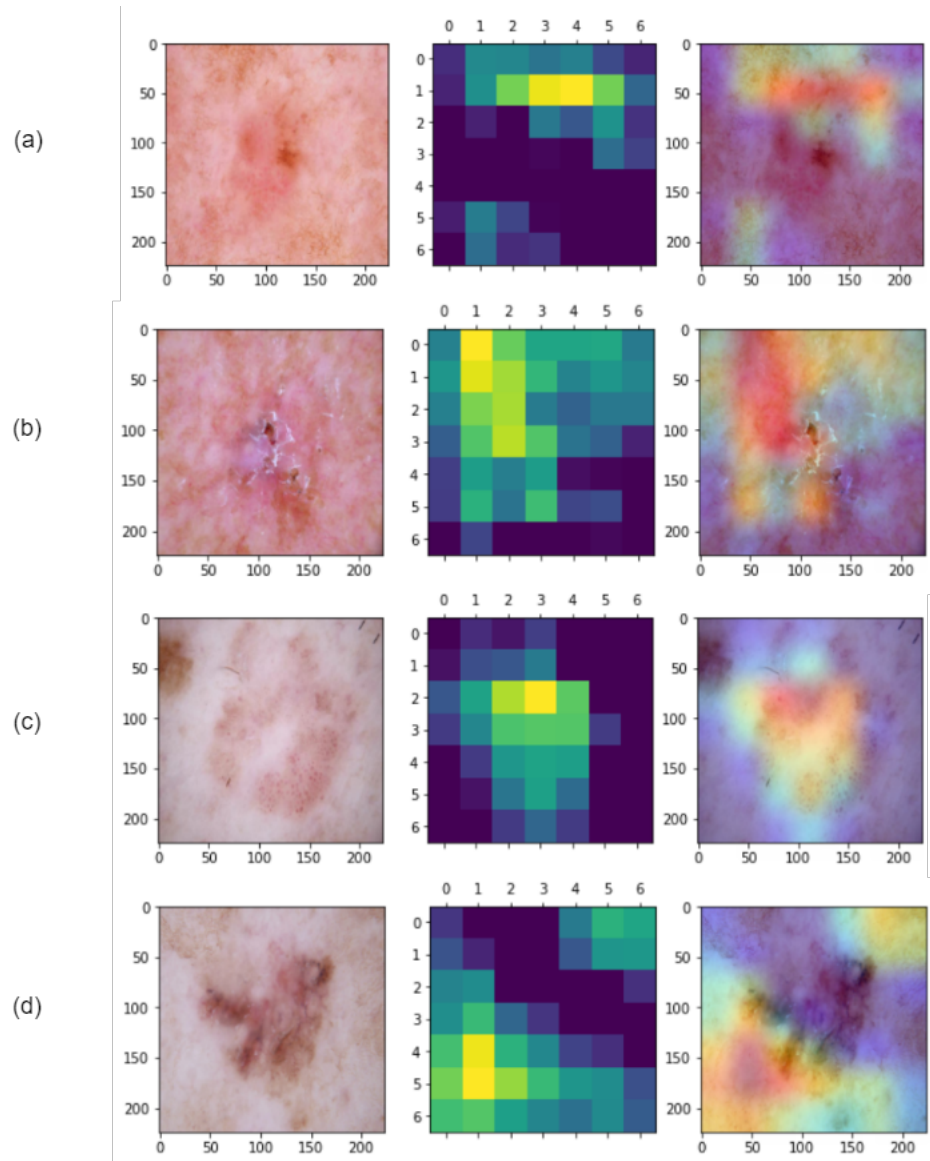
FIGURE 6.14: Visual explanations of incorrect classifications of the HAM10000 and BCN20000 data set via gradient based localization. First column is the original image, second column represents the heatmap and third column is the superimposed heatmap on the image. (a) AK classified as SCC. (b) AK classified as BCC. (c) AK classified as BCC. (d) SCC classified as BCC.

The accuracy of diagnosis with the HAM10000 data set was lower than that with the BCN20000 data set, possibly due to differences in the size of the data set, patient ethnicity, and general contrast variations in the images. A bigger data set was created with a combination of the two, and used for training, which obtained better results than the two alone, 82% in testing compared to 76% and 79% respectively. However, in all the experiments presented, there is a common point, as can be seen in the summary Tables 6.3, 6.4 and 6.5 there is overfitting in most cases. Models are learning to fit the data only to the training set, which would explain the low versatility of the models to be used with another data set. When tested with other data sets, the most versatile
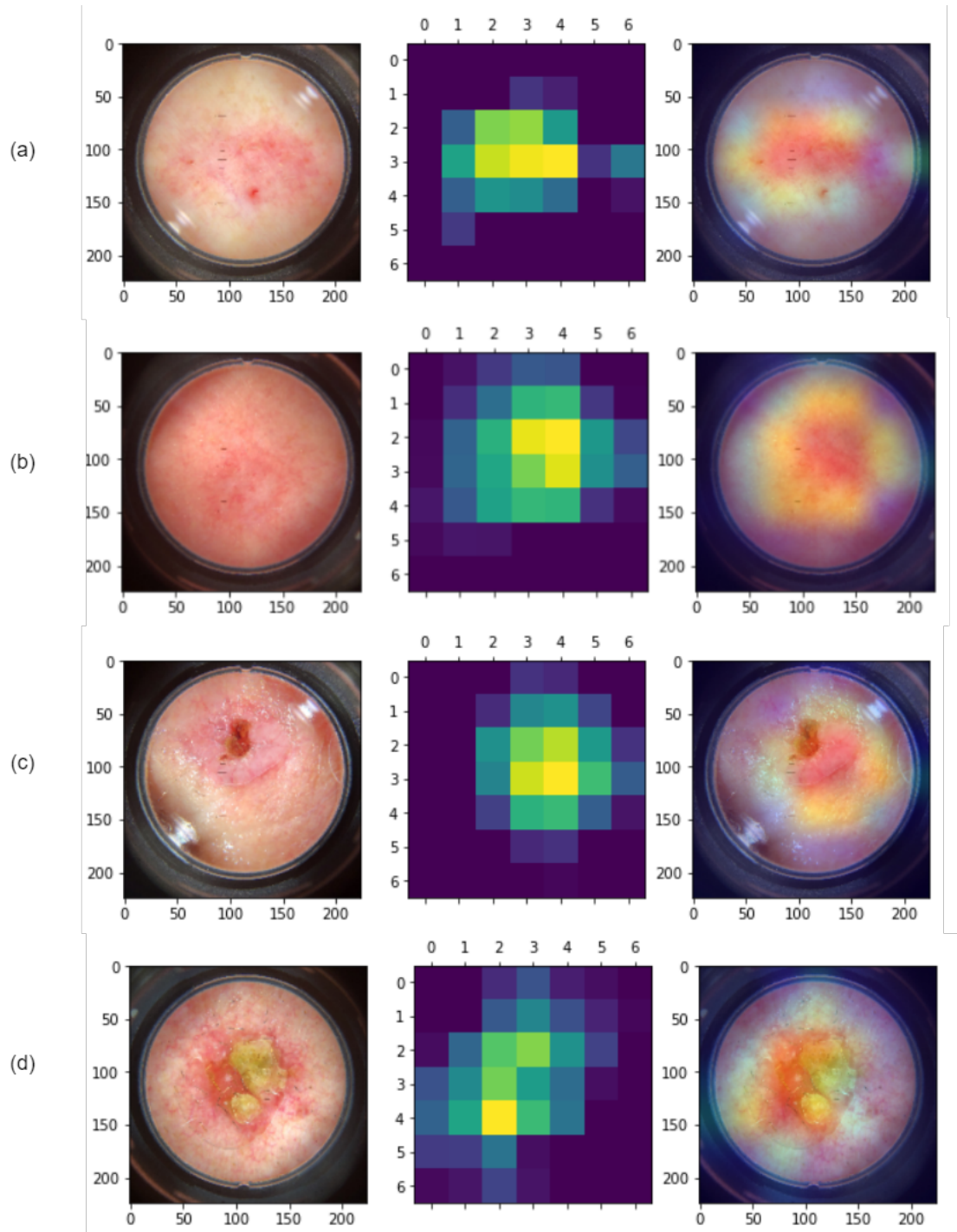
FIGURE 6.15: Visual explanations of correct classifications of the LU200 data set via gradient based localization. First column is the original image, second column represents the heatmap and third column is the superimposed heatmap on the image. (a) BCC. (b) BCC. (c) BCC. (d) SCC.

model was the one trained with HAM10000 and BCN20000 combined, which obtained 65% and 73% respectively and 66% accuracy in a completely new data set as LU200. However for LU200, despite the good accuracy obtained, when looking at the confusion matrix, the model classified most cases as BCC and since the LU200 is unbalanced with
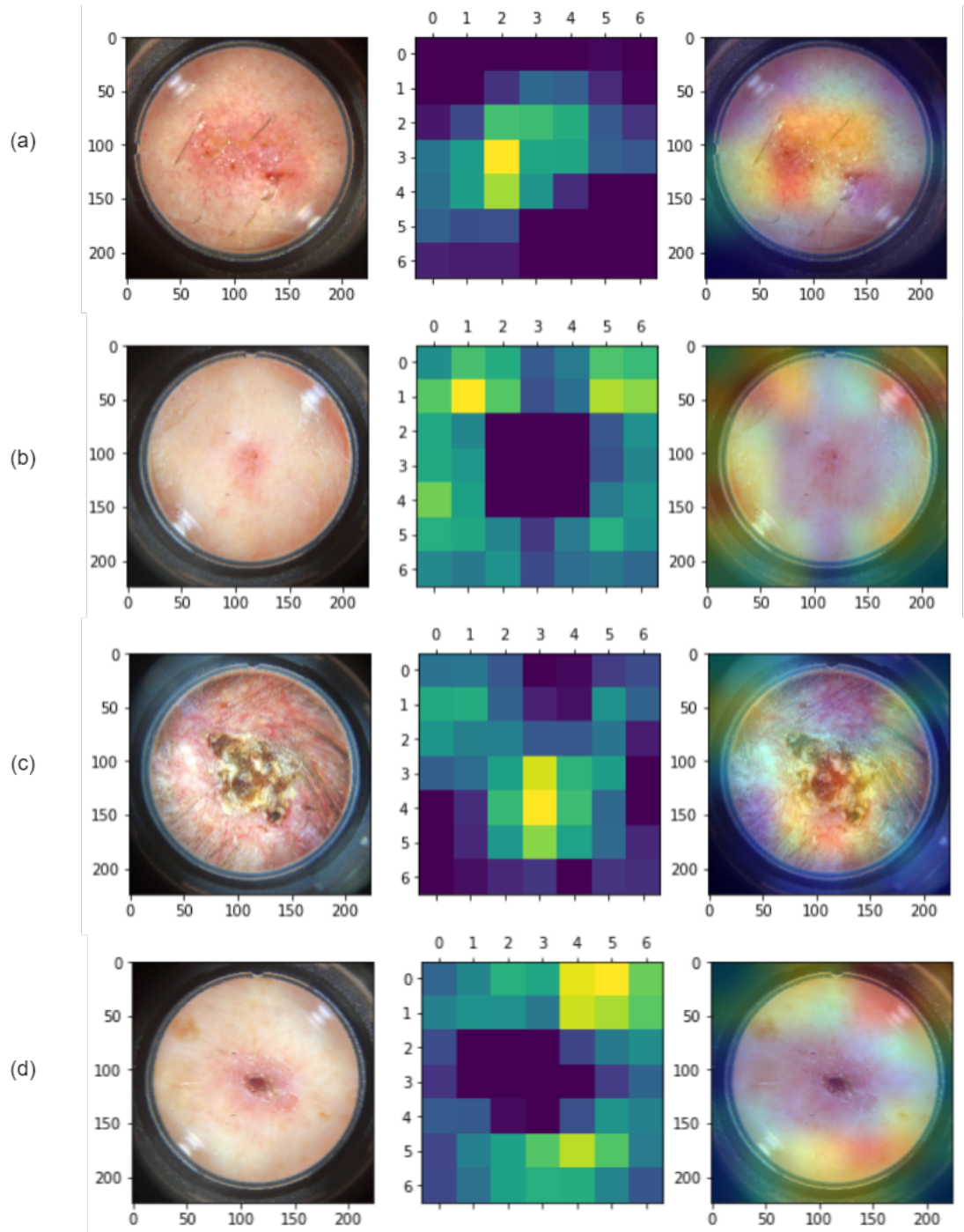
FIGURE 6.16: Visual explanations of wrong classifications of the LU200 data set via gradient based localization. First column is the original image, second column represents the heatmap and third column is the superimposed heatmap on the image. (a) AK classified as BCC. (b) AK classified as BCC (c) BCC classified as SCC. (d) SCC classified as BCC.

more BCC cases the accuracy turns high when in reality only 20% of the AK and 13% of the SCC cases were correctly classified.

To train our models in the experiments we used transfer learning, the models used were
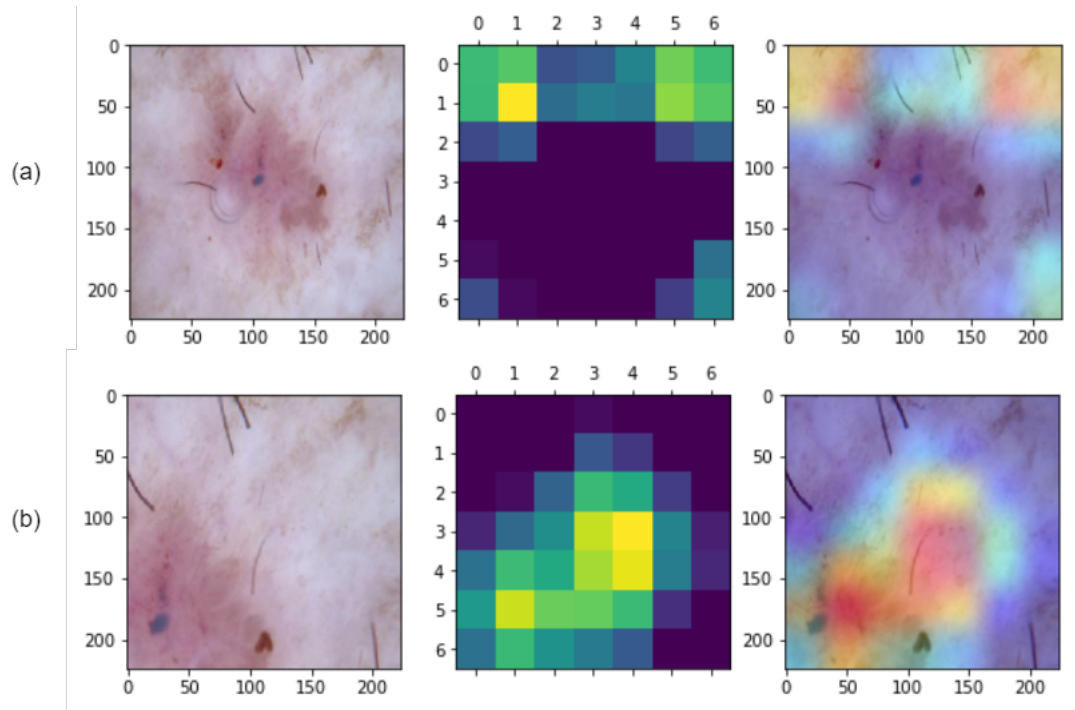
FIGURE 6.17: Visual explanations of correct classifications of the HAM10000 and BCN20000 data set via gradient based localization (a) BCC classified as BCC. (b) BCC classified as BCC.
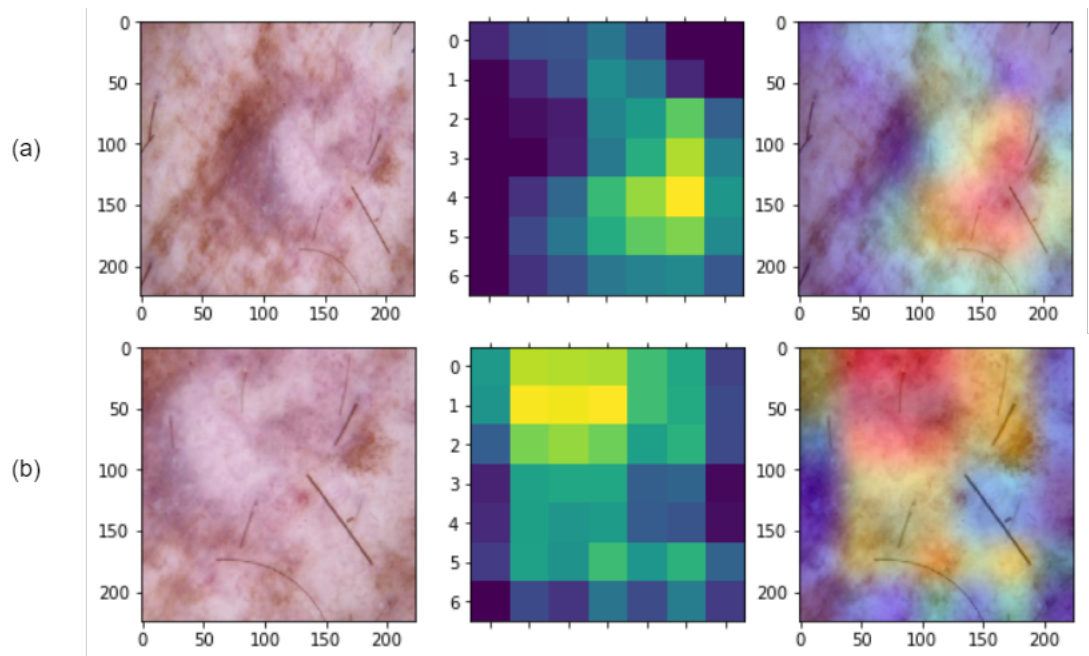
FIGURE 6.18: Visual explanations of wrong classifications of the HAM10000 and BCN20000 data set via gradient based localization (a) AK classified as AK. (b) AK classified as BCC.

InceptionResNetV2, DenseNet and Mobilenet. The best models were obtained from DenseNet which have proven to be good for image classification and can effectively solve the gradient vanishing problem, improve feature map propagation, and minimise

some parameters. Although Mobilenet introduces blocks similar to DenseNet, its small architecture and size, perfect for an on-device application, does not achieve as good accuracy as larger models.

Chapter 7

Conclusions

In this chapter, the conclusions of this master thesis project are presented, as well as the completion of the objectives and the future work of the project.

7.1 Conclusions

The objective of this project was to develop a deep learning algorithm to help dermatologists classify non-pigmented skin lesions from dermatocopical images. The results of the development are the following:

- **Research of skin cancer data sets available.** A thorough analysis was carried out to find the most available online data sets with their corresponding characteristics, the full analysis is presented in the Appendix A. The decision of which data sets to use was made based on the skin type, if it could be similar to the Swedish type, if the images were dermatoscopic, if our target non-pigmented lesions were included, and the amount of pictures. The data sets that met this criteria were the HAM10000 and BCN20000 data sets.

- **Analysis and evaluation of different deep learning model architectures.** The experiments in the project were carried out using the transfer learning technique. It was found that the best approach was to retrain an existing network with a similar task of image classification to solve our problem of a small data set. The models used were InceptionResNetV2, DenseNet and Mobilenet. In the experiments carried out through the project, DenseNet was the network from which the best results were obtained. Although Mobilenet did not achieve as good scores as DenseNet, its compact architecture and size make it perfect for an on-device

application, thus it is a good consideration depending where the model will be running once it is deployed.

- **Analysis and experimental validation of the proposed solution.** The experiments were carried out using the architectures proposed and available in Keras for transfer learning. For training, we used the data sets previously selected and although they were unbalanced with a higher number of BCC images and without a preprocessing step applied to them in this project, the experimental results are promising and give hope for a possible future solution with good predictive capabilities that will improve the quality of the diagnosis made in dermatological clinics. The results obtained are different for each of the data sets used, HAM10000 with a test accuracy of 76%, BCN20000 with 79% and both together 82%. It is important to consider that we are in a medical environment in which the precision of the predictions need to meet the standards since there are lives at stake.

- **Analysis whether the type of skin affects the classification.** The versatility of trained networks was tested with another data set different from the training one. The models performed poorly when used in another data set, due to differences in the images from one to another, such as image device and the skin type surrounding the lesion. Whether it is the type of image or the type of skin, it is not possible to tell from the tests performed. It is necessary to remove the bias in the image (e.g. black circle, glares, etc.) to focus on the skin type. Although the way the lesions present themselves is different depending on the skin type, it has been demonstrated in other publications that it is a conditioning factor when testing the versatility of a network [28, 76].

- **Analysis of the key elements on the image that lead to diagnosis.** The Grad-Cam algorithm was implemented and explained in Section 6.4 to determine in which area of the image the model was focussing to base its predictions. The model used was the combination of HAM10000 and BCN20000, which revealed that the model focused on the lesion but in most cases it was also focussing on the "white clues" (such as white circles, white lines, white structureless areas), in which also dermatologists focus when making a diagnosis for non-pigmented skin lesions.

Finally, we can conclude that the objectives specified at the beginning of the project have been met at the end of the project.

7.2   Future work

In this section, we present the possible future lines of the project:

- **Correction of image prior to training.** Add a correction before inserting the image into the model, perform filtering and cropping to normalise the images, such as removing the black border of the camera or correcting present reflections.

- **Optimization of the model for better performance.** A further study and more experiments that include a more variety of images and modify more hyper-parameters will allow us to obtain more robust models, in terms of prediction accuracy and possible reducing the current overfitting.

- **Adding prediction for the following procedure to treat the lesion.** In the new Lund hospital, LU200 data set, information is collected about the procedure selected to treat the lesion. Using this information for training, the model could provide feedback to the doctor after the prediction with the procedure preferred for that lesion. Furthermore, more information could be added, such as the certainty of prediction and the hot spots of the lesion to look at.

- **Expand the databases.** Continue with the data collection at the Skåne University Hospital in Lund, to allow for a better model prediction with consistent data taken with the same device and skin type.

- **Add corrective capability.** It would be desirable to incorporate the ability to correct the model decisions by a doctor to automatically gather feedback that can allow us to refine the behaviour of the model.

- **Re-training model while it is deployed.** Once the model is deployed for production and in use, how the lesions appear could evolve. Monitoring the new data and using these one would allow the model to be adapted for the current necessities.

Bibliography

[1] World Health Organization. Estimated number of new cases from 2020 to 2040 in sweden, . URL `https://gco.iarc.fr/tomorrow/en/dataviz/trends?types=0&sexes=1_2&mode=cancer&group_populations=1&multiple_populations=1&multiple_cancers=1&cancers=17_16&populations=752`.

[2] International Skin Imaging Collaboration. Isic archive. URL `https://www.isic-archive.com/`.

[3] DermNe NZ. Rosettes, 2019. URL `https://dermnetnz.org/topics/rosettes`.

[4] IBM Cloud Education. Neural networks, 2020. URL `https://www.ibm.com/cloud/learn/neural-networks`.

[5] World Cancer Reasearch Fund. Skin cancer statistics. URL `https://www.wcrf.org/dietandcancer/skin-cancer-statistics/`.

[6] World Health Organization. Global cancer observatory, . URL `https://gco.iarc.fr/`.

[7] Amirreza Rezvantalab, Habib Safigholi, and Somayeh Karimijeshni. Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. *CoRR*, abs/1810.10348, 2018. URL `http://arxiv.org/abs/1810.10348`.

[8] M Manoj krishna, M Neelima, M Harshali, and M Venu Gopala Rao. Image classification using deep learning. *Int. j. eng. technol.*, 7(2.7):614, March 2018.

[9] R C Eberhart and R W Dobbins. Early neural network development history: the age of camelot. *IEEE Eng. Med. Biol. Mag.*, 9(3):15–18, 1990.

[10] Lijuan Liu, Yanping Wang, and Wanle Chi. Image recognition technology based on machine learning. *IEEE Access*, pages 1–1, 2020.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[13] Anirudha Ghosh, Abu Sufian, Farhana Sultana, Amlan Chakrabarti, and Debashis De. Fundamental concepts of convolutional neural network. In *Intelligent Systems Reference Library*, Intelligent systems reference library, pages 519–567. Springer International Publishing, Cham, 2020.

[14] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993.

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[16] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. URL http://arxiv.org/abs/1602.07261.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[20] Muralikrishna Puttagunta and S Ravi. Medical image analysis based on deep learning approach. *Multimed. Tools Appl.*, 80(16):24365–24398, April 2021.

[21] Philipp Tschandl, Cliff Rosendahl, Bengu Nisa Akay, Giuseppe Argenziano, Andreas Blum, Ralph P Braun, Horacio Cabo, Jean-Yves Gourhant, Jürgen Kreusch, Aimilios Lallas, Jan Lapins, Ashfaq Marghoob, Scott Menzies, Nina Maria Neuber, John Paoli, Harold S Rabinovitz, Christoph Rinner, Alon Scope, H Peter Soyer, Christoph Sinz, Luc Thomas, Iris Zalaudek, and Harald Kittler. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol.*, 155(1):58–65, January 2019.

[22] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D. Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, Edmund Yick Mun Wong, Charumathi Sabanayagam, Mani

Baskaran, Farah Ibrahim, Ngiap Chuan Tan, Eric A. Finkelstein, Ecosse L. Lamoureux, Ian Y. Wong, Neil M. Bressler, Sobha Sivaprasad, Rohit Varma, Jost B. Jonas, Ming Guang He, Ching-Yu Cheng, Gemmy Chui Ming Cheung, Tin Aung, Wynne Hsu, Mong Li Lee, and Tien Yin Wong. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, 318 (22):2211–2223, 12 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.18152. URL `https://doi.org/10.1001/jama.2017.18152`.

[23] Justin Ker, Yeqi Bai, Hwei Yee Lee, Jai Rao, and Lipo Wang. Automated brain histology classification using machine learning. *Journal of Clinical Neuroscience*, 66:239–245, August 2019. doi: 10.1016/j.jocn.2019.05.019. URL `https://doi.org/10.1016/j.jocn.2019.05.019`.

[24] Fabio A. Spanhol, Luiz S. Oliveira, Paulo R. Cavalin, Caroline Petitjean, and Laurent Heutte. Deep features for breast cancer histopathological image classification. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1868–1873, 2017. doi: 10.1109/SMC.2017.8122889.

[25] Long Gao, Lei Zhang, Chang Liu, and Shandong Wu. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artificial Intelligence in Medicine*, 108:101935, August 2020. doi: 10.1016/j.artmed.2020.101935. URL `https://doi.org/10.1016/j.artmed.2020.101935`.

[26] Nih/ncats workshop – "machine intelligence in healthcare: Perspectives on trustworthiness, explainability, usability and transparency". URL `https://www.acadrad.org/event/nih-ncats-workshop-machine-intelligence-in-healthcare-perspectives-on-trustworthiness-explainability-usability-and-transparency/`.

[27] Christine M. Cutillo, , Karlie R. Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, and Kenneth D. Mandl. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *npj Digital Medicine*, 3(1), March 2020. doi: 10.1038/s41746-020-0254-2. URL `https://doi.org/10.1038/s41746-020-0254-2`.

[28] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, and Rubeta N Matin. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, January 2022. doi: 10.1016/s2589-7500(21)00252-1. URL `https://doi.org/10.1016/s2589-7500(21)00252-1`.

[29] Cliff Rosendahl, Alan Cameron, Philipp Tschandl, Agata Bulinska, Iris Zalaudek, and Harald Kittler. Prediction without pigment: a decision algorithm for non-pigmented skin malignancy. *Dermatology Practical & Conceptual*, January 2014. doi: 10.5826/dpc.0401a09. URL `https://doi.org/10.5826/dpc.0401a09`.

[30] Instituto Nacional del cancer. Lunares comunes, nevos displásicos y el riesgo de melanoma-instituto nacional del cáncer., 2019. URL `https://www.cancer.gov/espanol/tipos/piel/hoja-informativa-lunares`.

[31] Mehwish Dildar, Shumaila Akram, Muhammad Irfan, Hikmat Ullah Khan, Muhammad Ramzan, Abdur Rehman Mahmood, Soliman Ayed Alsaiari, Abdul Hakeem M Saeed, Mohammed Olaythah Alraddadi, and Mater Hussen Mahnashi. Skin cancer detection: A review using deep learning techniques. *Int. J. Environ. Res. Public Health*, 18(10):5479, May 2021.

[32] Simon Kalouche. Vision-based classification of skin cancer using deep learning. 2016.

[33] Aya Abu Ali and Hasan Al-Marzouqi. Melanoma detection using regular convolutional neural networks. In *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*. IEEE, November 2017.

[34] Chengxi Ye, Chen Zhao, Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. Lightnet: A versatile, standalone matlab-based environment for deep learning. *CoRR*, abs/1605.02766, 2016. URL `http://arxiv.org/abs/1605.02766`.

[35] Mahamudul Hasan, Surajit Das Barman, Samia Islam, and Ahmed Wasif Reza. Skin cancer detection using convolutional neural network. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence - ICCAI '19*, New York, New York, USA, 2019. ACM Press.

[36] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017. doi: 10.1109/TMI.2016.2642839.

[37] Iris Zalaudek, Jürgen Kreusch, Jason Giacomel, Gerardo Ferrara, Caterina Catricalà, and Giuseppe Argenziano. How to diagnose nonpigmented skin tumors: A review of vascular structures seen with dermoscopy. *Journal of the American Academy of Dermatology*, 63(3):361–374, September 2010. doi: 10.1016/j.jaad.2009.11.698. URL `https://doi.org/10.1016/j.jaad.2009.11.698`.

[38] Ulzii-Orshikh Dorj, Keun-Kwang Lee, Jae-Young Choi, and Malrey Lee. The skin cancer classification using deep convolutional neural network. *Multimed. Tools Appl.*, 77(8):9909–9924, April 2018.

[39] Hon A/Prof Amanda Oakley. Fitzpatrick skin phototype, by dermnetnz, 2012. URL `https://dermnetnz.org/topics/skin-phototype`.

[40] Dr Amanda Oakley. Dermoscopy, by dermnetnz, 2004. URL `https://dermnetnz.org/topics/dermoscopy`.

[41] Dermoscopedia. Non polarized dermoscopy, 2020. URL `https://dermoscopedia.org/Non_polarized_dermoscopy`.

[42] Dermoscopedia. Seborrheic keratoses, 2019. URL `https://dermoscopedia.org/Seborrheic_keratoses`.

[43] DermNe NZ. Melanocytic naevi pathology, 2013. URL `https://dermnetnz.org/topics/melanocytic-naevi-pathology`.

[44] DermNe NZ. Dermoscopy of benign melanocytic lesions, 2008. URL `https://dermnetnz.org/cme/dermoscopy-course/dermoscopy-of-benign-melanocytic-lesions#:~:text=Specific%20dermatoscopic%20features%20suggesting%20melanocytic,steel%2Dblue%20of%20blue%20naevus`.

[45] DermNe NZ. Dermatofibroma, 2016, Revised 2020. URL `https://dermnetnz.org/topics/dermatofibroma`.

[46] DermNe NZ. Dermoscopy of melanoma, 2008. URL `https://dermnetnz.org/cme/dermoscopy-course/dermoscopy-of-melanoma`.

[47] Dermoscopedia. Actinic keratosis, 2019. URL `https://dermoscopedia.org/Actinic_keratosis`.

[48] Dermoscopedia. Basal cell carcinoma, 2019. URL `https://dermoscopedia.org/Basal_cell_carcinoma`.

[49] Junji Kato, Kohei Horimoto, Sayuri Sato, Tomoyuki Minowa, and Hisashi Uhara. Dermoscopy of melanoma and non-melanoma skin cancers. *Frontiers in Medicine*, 6, 2019. ISSN 2296-858X. doi: 10.3389/fmed.2019.00180. URL `https://www.frontiersin.org/article/10.3389/fmed.2019.00180`.

[50] Dermoscopedia. Squamous cell carcinoma, 2020. URL `https://dermoscopedia.org/Squamous_cell_carcinoma`.

[51] Olabe, X. B. (s.f.). Redes neuronales artificiales y sus aplicaciones, 2014.

[52] Google Developers. Machine learning glossary. URL `https://developers.google.com/machine-learning/glossary/`.

[53] Keras Documentation. What does "sample", "batch", "epoch" mean?, . URL `https://keras.io/getting-started/faq/#what-does-sample-batch-epoch-mean`.

[54] Jason Brownlee . What is the difference between a batch and an epoch in a neural network?, 2018. URL `https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/`.

[55] Jason Brownlee. How to configure the learning rate hyperparameter when training deep learning neural networks, 2019. URL `https://machinelearningmastery.com/learning-rate-for-deep-learning-neural-networks/`.

[56] Suki Lau. Learning rate schedules and adaptive learning rate methods for deep learning, 2017. URL `https://towardsdatascience.com/learning-rate-schedules-and-adaptive-learning-rate-methods-for-deep-learning-2c8f433990d1`.

[57] Carolina Bento. Stochastic gradient descent explained in real life, 2021. URL `https://towardsdatascience.com/stochastic-gradient-descent-explained-in-real-life-predicting-your-pizzas-cooking-time-b7639d5e6a32`.

[58] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[59] Sagar Sharma. Activation functions in neural networks, 2017. URL `https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6`.

[60] MissingLink.ai. 7 types of neural network activation functions: How to choose? URL `https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/`.

[61] Ravindra Parmar. Common loss functions in machine learning. URL `https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23`.

[62] Raúl Gómez. Understanding categorical cross-entropy loss, binary cross-entropy loss, softmax loss, logistic loss, focal loss and all those confusing names. URL `https://gombru.github.io/2018/05/23/cross_entropy_loss/`.

[63] Scikit-learn documentation. Scikit-learn, machine learning in python. URL `https://scikit-learn.org/stable/index.html`.

[64] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.191.

[65] Keras Documentation. Keras applications, avaliable models, . URL `https://keras.io/applications/`.

[66] Keras. Keras api. URL `https://keras.io/`.

[67] Jupyter. Jupyter documentation. URL `https://jupyter.org/`.

[68] Anaconda. Anaconda official webpage. URL `https://www.anaconda.com/`.

[69] Dermicus. Dermicus platform, . URL `https://dermicus.com/en/platform/`.

[70] Dermicus. Heine ic1, . URL `https://dermicus.com/dermatoskop/heine-ic1/`.

[71] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), Aug 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018. 161. URL `http://dx.doi.org/10.1038/sdata.2018.161`.

[72] ISIC. Isic-2018 challenge, 2018. URL `https://challenge.isic-archive.com/landing/2018/`.

[73] ISIC. Isic-2019 challenge, 2019. URL `https://challenge2019.isic-archive.com/`.

[74] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. Bcn20000: Dermoscopic lesions in the wild, 2019. URL `https://arxiv.org/abs/1908.02288`.

[75] fChollet. Grad-cam class activation visualization, 2021. URL `https://keras.io/examples/vision/grad_cam/`.

[76] Seung Seog Han, Myoung Shin Kim, Woohyung Lim, Gyeong Hun Park, Ilwoo Park, and Sung Eun Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J. Invest. Dermatol.*, 138(7):1529–1538, July 2018.

Appendix A

Publicly available data sets

In this chapter the research on publicly available data sets is presented in the form of a table.

Name	Country	Year	Image modality	Image acquisition device	Skin lesions included	Number of participants	Number of images
Asan	South Korea	2017	Macroscopic	Not reported	Basal cell carcinoma 1,082 Squamous cell carcinoma 1,231 Intraepithelial carcinoma 918 Actinic keratosis 651 Seborrheic keratosis 1,423 Malignant melanoma 599 Melanocytic nevus 2,706 Lentigo 1,193 Pyogenic granuloma Hemangioma 2,715 Dermatofibroma 1,247 Wart 2,985	4867	17125
Hallym	South Korea	2017	Not reported	Not reported	Basal cell carcinoma 152	106	152
SNU dataset entire test set	South Korea	2018	Macroscopic	Not reported	134 disorders (5 malignancies and 129 non-malignancies)	1608	2201
SNU testset	South Korea	2018	Macroscopic	Not reported	134 disorders (5 malignancies and 129 non-malignancies)	Not reported	240
Severance dataset (test subset)	South Korea	2020	Macroscopic	Not reported	More than 43 disorders, some relevant ones: Basal cell carcinoma 643 (6.3%), Squamous cell carcinoma 158 (1.5%), Intraepithelial carcinoma (SCC in situ) 255 (2.5%), Actinic keratosis 784 (7.6%)	10426	40331
PH2	Portugal	2013	Dermoscopy	Tuebinger Mole Analyzer System	MEL 40 NV 160	Not reported	200
SKINL2	Portugal	2019	Light field photographs, dermoscopic photographs (paired)	Raytrix R42 Galilean focused plenoptic camera, Ricoh 25 mm	Melanoma Melanocytic Nevus Basal Cell Carcinoma Seborrheic Keratosis Hemangioma Dermatofibroma Psoriasis Other	Not reported	814
ISIC gallery			Clinical, Dermoscopy, Unknown		Basal Cell Carcinoma 3396, Squamous Cell Carcinoma 656, etc		
ISIC 2020 Hospital Clinic Barcelona	Spain	2020	Dermoscopic	MoleMax HD digital dermatoscopy system	Melanoma 179, Unknown 7132	356	7311
BCN20000	Spain	2019	Dermoscopic	Dermoscopic attachments to three high-resolution cameras	Melanoma 2700 Melanocytic nevus 4100 Basal cell carcinoma 2700 Actinic keratosis 600 Dermatofibroma 100, Vascular lesion 100, Squamous cell carcinoma 500 Serborheic kerastosis 1000	Not reported	19424
ISIC 2020 University of Queensland	Australia	2020	Desmoscopic	Not reported	Not reported	304	8449
ISIC 2020 Sydney Melanoma Diagnosis Centre and Melanoma Institute Australia	Australia	2020	Dermoscopic	Dermoscopic attachment to a digital single reflex lens camera or a smartphone	Atypical melanocytic proliferation 1, Letingo NOS 6, Lichenoid kerastosis 3, Melanoma 134, Nevus 161, Seborrheic kerastosis 11, Solar Letingo 4, Unknown 1563	441	1884

Dataset	Country	Year	Modality	Device	Lesion classes		Patients	Images
ISIC 2020 Medical University Vienna	Austria	2020	Dermoscopic	MoleMax HD digital dermatoscopy system	Melanoma 43, Nevus 4331		432	4374
HAM10000	Austria and Australia	2018	Demoscopic	Various devices including: MoleMax HD, DermLite Foto (3Gen) camera, DermLite Fluid, DermLite DL3, and analogue cameras	BCC 514 BKL 1099 AKIEC 327 DF 115 MEL 1113 NV 6705 VASC 142		Not reported	10015
ISIC 2020 Memorial Sloan Ketering Cancer Centre	USA	2020	Dermoscopic	Dermoscopic attachment to a digital single reflex lens camera or a smartphone	Letingo NOS 34, Lichenoid kerastosis 28, Melanoma 216, Nevus 614, Unknown 10111		523	11108
PAD-UFES -20	Brazil	2020	Macroscopic	Smartphones	Actinic Keratosis (ACK) 543 Basal Cell Carcinoma (BCC) 442 Melanoma (MEL) 67 Nevus (NEV) 196 Squamous Cell Carcinoma (SCC) 149 Seborrheic Keratosis (SEK) 215		1373	2298
MED-NODE	Netherlands	2015	Macroscopic	Nikon D3 or Nikon D1x body and a Nikkor 2·8/105 mm micro lens	Melanoma 70, Nevus 100		Not reported	170
MSK 1–5	Not reported	MSK 1-2 2015 MSK 3-5 2017	Demoscopic	Not reported	Actinic Keratosis 2, Angiofibroma 1, Angioma 13, Atypical melanocytic proliferation 13, Basal Cell Carcinoma 30, Dermatofibroma 5, Letingo NOS 71, Letingo simplex 27, Melanoma 826, Nevus 2202, Squamous Cell Carcinoma 5, Seborrheic Keratosis 413, Scar 1, Solar letingo 57, Unknown 242		Not reported	3918
UDA1-2	Not reported	UDA-1 2014 UDA-2 2015	Demoscopic	Not reported	Basal Cell Carcinoma 3, Dermatofibroma 2, Lichenoid kerastosis 1, Melanoma 193, Nevus 408, Seborrheic kerastosis 6, Angioma 2, Unknown 2		Not reported	617
University of Waterloo dataset	Not reported	Not reported	Macroscopic	Consumer level cameras		2	Not reported	206
SD-260	Not reported	2019	Macroscopic	Digital cameras and mobile phones	260 categories		Not reported	20600
IMA 205	Not reported	2018	Not reported	Not reported	Only binary malignant or bening		Not reported	900
7-point criteria evaluation database	Not reported	2018	Dermoscopic and	Not reported	Basal Cell Carcinoma 42, Blue Nevus 28 , Clark nevus 399, Combined nevus 13, Congenital nevus 17, Dermal nevus 33, Recurrent nevus 6, Reed or Spitz nevus 79, Melanoma 252, Dermatofibroma 20, Letingo 24, Melanosis 16, Miscellaneous 8, Vascular lesion 29, Seborreheic kerastosis 45		1011	2013
2018 JID editorial images	Caucasian patients in the southern United States	2018	Macroscopic	Not reported	Melanoma 37, Basal Cell Carcinoma 40, Squamous Cell Carcinoma 23		Not reported	100
Atlas	United States.	Not reported	Not reported	Not reported	Basal Cell Carcinoma 1561, Seborrheic kerastosis 897, Malignant Melanoma 228, Melanotyc nevus 626, Letingo 73, Others 473		Not reported	3820
Papadakis et al (2021) dataset	Germany	2021	Macroscopic	Commercial digital camera		1	156	156
MoleMapper app patient photos	USA	2017	Macroscopic	iphone (4s or newer)		2	2069	2422
DERMOFIT Image Library: Edinburgh dataset	UK	Not reported	Not reported	Canon EOS 350D DSLR	Basal Cell Carcinoma 239, Squamous Cell Carcinoma 88, Intrapithelial Carcinoma 78, Actinic Kerastosis 45, Seborreheic kerastosis 257, Malignant Melanoma 76, Melanotyc nevus 331, Pyogenic granuloma 24, Hemangioma 97, Dermatofibroma 65		Not reported	1300

Appendix B

Architectures from the deep learning experiments

In this chapter, the different architectures used to train the various experiments from section 6.2 are presented, following the same previous order.

```
Layer (type)                    Output Shape            Param #
=================================================================
inception_resnet_v2 (Functio   (None, 1536)            54336736

dense_4 (Dense)                 (None, 512)             786944

dropout_3 (Dropout)             (None, 512)             0

dense_5 (Dense)                 (None, 256)             131328

dropout_4 (Dropout)             (None, 256)             0

dense_6 (Dense)                 (None, 128)             32896

dropout_5 (Dropout)             (None, 128)             0

flatten_1 (Flatten)             (None, 128)             0

dense_7 (Dense)                 (None, 3)               387
=================================================================
Total params: 55,288,291
Trainable params: 55,227,747
Non-trainable params: 60,544
```

FIGURE B.1: Architecture from experiment 1 in the HAM10000 model.

```
Layer (type)                    Output Shape          Param #
=================================================================
densenet201 (Functional)        (None, 7, 7, 1920)    18321984

global_average_pooling2d_4 (    (None, 1920)          0

dense_11 (Dense)                (None, 256)           491776

dropout_7 (Dropout)             (None, 256)           0

dense_12 (Dense)                (None, 128)           32896

dropout_8 (Dropout)             (None, 128)           0

dense_13 (Dense)                (None, 3)             387
=================================================================
Total params: 18,847,043
Trainable params: 18,617,987
Non-trainable params: 229,056
```

FIGURE B.2: Architecture from experiment 2 in the HAM10000 model.

```
Layer (type)                    Output Shape          Param #
=================================================================
mobilenetv2_1.00_224 (Functi    (None, 7, 7, 1280)    2257984

global_average_pooling2d_3 (    (None, 1280)          0

dense_6 (Dense)                 (None, 1024)          1311744

dropout_4 (Dropout)             (None, 1024)          0

dense_7 (Dense)                 (None, 512)           524800

dropout_5 (Dropout)             (None, 512)           0

dense_8 (Dense)                 (None, 256)           131328

dropout_6 (Dropout)             (None, 256)           0

flatten_2 (Flatten)             (None, 256)           0

dense_9 (Dense)                 (None, 3)             771
=================================================================
Total params: 4,226,627
Trainable params: 1,968,643
Non-trainable params: 2,257,984
```

FIGURE B.3: Architecture from experiment 3 in the HAM10000 model.

```
Layer (type)                  Output Shape              Param #
=================================================================
densenet201 (Functional)      (None, 7, 7, 1920)        18321984
_____
global_average_pooling2d_3 (  (None, 1920)              0
_____
flatten_3 (Flatten)           (None, 1920)              0
_____
dense_3 (Dense)               (None, 3)                 5763
=================================================================
Total params: 18,327,747
Trainable params: 292,163
Non-trainable params: 18,035,584
```

FIGURE B.4: Architecture from experiment 1 in the BCN20000 model.

```
Layer (type)                  Output Shape              Param #
=================================================================
densenet201 (Functional)      (None, 1920)              18321984
_____
dense (Dense)                 (None, 1024)              1967104
_____
dropout (Dropout)             (None, 1024)              0
_____
dense_1 (Dense)               (None, 512)               524800
_____
dropout_1 (Dropout)           (None, 512)               0
_____
dense_2 (Dense)               (None, 256)               131328
_____
dropout_2 (Dropout)           (None, 256)               0
_____
dense_3 (Dense)               (None, 3)                 771
=================================================================
Total params: 20,945,987
Trainable params: 2,624,003
Non-trainable params: 18,321,984
```

FIGURE B.5: Architecture from experiment 2 in the BCN20000 model.

```
Layer (type)                  Output Shape              Param #
=================================================================
mobilenetv2_1.00_224 (Functi  (None, 7, 7, 1280)        2257984
_____
global_average_pooling2d (Gl  (None, 1280)              0
_____
flatten (Flatten)             (None, 1280)              0
_____
dense (Dense)                 (None, 3)                 3843
=================================================================
Total params: 2,261,827
Trainable params: 3,843
Non-trainable params: 2,257,984
```

FIGURE B.6: Architecture from experiment 3 in the BCN20000 model.

```
Layer (type)                    Output Shape            Param #
=================================================================
densenet201 (Functional)        (None, 7, 7, 1920)      18321984

global_average_pooling2d_2 (    (None, 1920)            0

flatten_2 (Flatten)             (None, 1920)            0

dense_5 (Dense)                 (None, 3)               5763
=================================================================
Total params: 18,327,747
Trainable params: 607,683
Non-trainable params: 17,720,064
```

FIGURE B.7: Architecture from experiment 1 in the HAM10000 and BCN20000 model.

```
Layer (type)                    Output Shape            Param #
=================================================================
densenet201 (Functional)        (None, 7, 7, 1920)      18321984

global_average_pooling2d_1 (    (None, 1920)            0

flatten_1 (Flatten)             (None, 1920)            0

dense_4 (Dense)                 (None, 3)               5763
=================================================================
Total params: 18,327,747
Trainable params: 292,163
Non-trainable params: 18,035,584
```

FIGURE B.8: Architecture from experiment 2 in the HAM10000 and BCN20000 model.

```
Layer (type)                    Output Shape            Param #
=================================================================
densenet201 (Functional)        (None, 7, 7, 1920)      18321984

global_average_pooling2d (Gl    (None, 1920)            0

dense (Dense)                   (None, 1024)            1967104

dropout (Dropout)               (None, 1024)            0

dense_1 (Dense)                 (None, 512)             524800

dropout_1 (Dropout)             (None, 512)             0

dense_2 (Dense)                 (None, 256)             131328

dropout_2 (Dropout)             (None, 256)             0

flatten (Flatten)               (None, 256)             0

dense_3 (Dense)                 (None, 3)               771
=================================================================
Total params: 20,945,987
Trainable params: 2,624,003
Non-trainable params: 18,321,984
```

FIGURE B.9: Architecture from experiment 3 in the HAM10000 and BCN20000 model.