Integration of heterogeneous data for the classiffication of non-melanoma skin cancer

Hoepel, F.M.C.

Award date:
2020

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Industrial Engineering & Innovation Sciences
Information Systems Group

Integration of Heterogeneous Data for the Classification of Non-Melanoma Skin Cancer

Master Thesis

16th November 2020

Frank M.C. Hoepel

Supervisors:
dr. Anna M. Wilbik (TU/e)
dr. Hendrik Eshuis (TU/e)
dr. Laura Genga (TU/e)
dr. Gertruud A.M. Krekels (MohsA)
drs. José D.van der Waa (MohsA)

Abstract

Skin cancer is the most upcoming cancer in modern societies. Non-melanoma skin cancers (NMSC) are the most common ones and require the most capacity and time in healthcare institutes. If patients can assess suspicious lesions themselves by making a picture and filling out a questionnaire, these patients can be consulted remotely and capacity can be saved. Therefore, this research is focused on the classification of NMSC using machine learning techniques that integrate both image data and numerical data. The thesis is twofold. Firstly, a convolutional neural network (CNN) is developed in order to make a classification of images of NMSC. Afterwards, this classification will serve as input for another model that integrates features that have been manually extracted from those images. Several models have been tested for this. The CNN reaches an accuracy of 83.4% in detecting NMSC, which can be improved with a random forest and logistic regression to 87.4%. Besides, the models are able to distinguish between different types of NMSC with relative high accuracy. These results show potential for application for both doctors and patients.

Executive summary

Introduction

Skin cancer is the most upcoming type of cancer in modern societies, especially in countries where a large portion of the population consists of fair-skinned people. A distinction can be made here between melanoma and non-melanoma skin cancer (NMSC). Melanoma are malignant and fast-growing skin cancers developing from melanocytes. Since melanoma are very severe and account for most of the deaths, most of the research has been done in that area, even though NMSC occur more often. Although NMSC are less severe and have lower mortality rates compared to the aggressive melanoma, a lot of the capacity at hospitals and dermatologists is spent on NMSC, which is very costly. Patients often demand immediate consultation about the spots on their skin, because they are afraid it could be immediately life-threatening. These consultations are executed on location and cost a lot of time and resources. This causes major planning problems.

A potential solution would be a method for patients to check for themselves whether they have skin cancer or not. This could mean that a patient takes a picture and enters some basic information about common risk factors for skin cancer, such as skin type or age, which would lead to a basic consult or risk estimation for the specific patient. Therefore, a desire exists for a model capturing this process of combining picture input and additional numerical information. Successful implementation of such a model would lead to significant labor reduction for hospitals and dermatologists and enhances long-term planning.

Hence, the reasons to write this thesis. The goal of this project is to build a model that combines the outputs of an image classification model with common features of NMSC in order to successfully classify images of NMSC. Over the late years, convolutional neural networks (CNN) have proven to be the superior method for image classification and will therefore be used to do the initial classification. Afterwards, a decision tree, a random forest and logistic regression will be used to evaluate whether that initial classification can be improved by the integration of features that have been extracted from the images.

Data & Methodology

The research consists of two parts, which is reflected in the structural design of this thesis project. Firstly, a CNN is designed that is able to distinct between 6 categories, which are actinic keratosis (AK), basal cell carcinoma (BCC) & squamous cell carcinoma (SCC) and 3 categories of lesions that are not skin cancers: Verrucae, Inflammatory Dermatoses & Naevi. The data set that will be used for this consists of 5308 images after the application of data augmentation techniques and will be tested on a test set of 277 images which have been kept apart. Afterwards, the prediction score of the neural network will, together with additional features retrieved from the images, serve as input for different methods that can handle heterogeneous data. The results of all the different models are evaluated on the same test set and related to the practical and theoretical relevance.

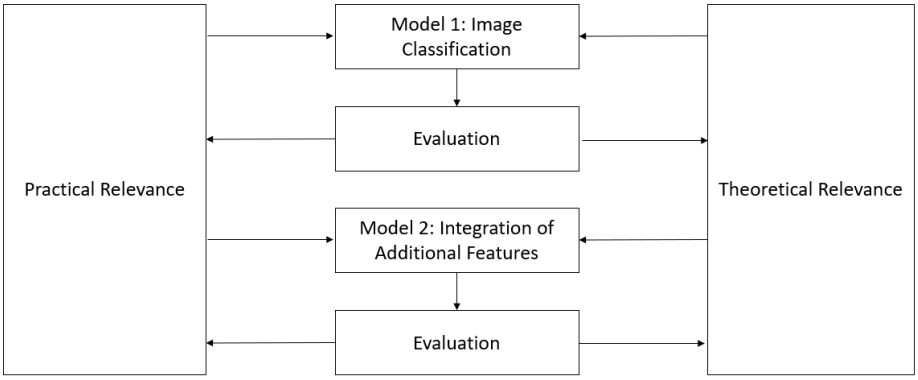

Figure 0.1: Design of the research structure

In the first part of the research consists of the development of a CNN. Two neural networks will be designed. The first one will be able to make a distinction between the 6 mentioned categories, and the second one will only make a distinction between skin cancer and no skin cancer.

The second part of the research includes the addition of features that have been extracted from the images of the different categories. These are features like location on the body, elevation, color & scaliness. Together with the output values from the convolutional neural networks, these features serve as input values for the feature integration methods, which leads to a final classification for the different techniques. This research structure is displayed in figure 0.2 and has been applied for the classification between the six different categories as well as the classification of NMSC in general.

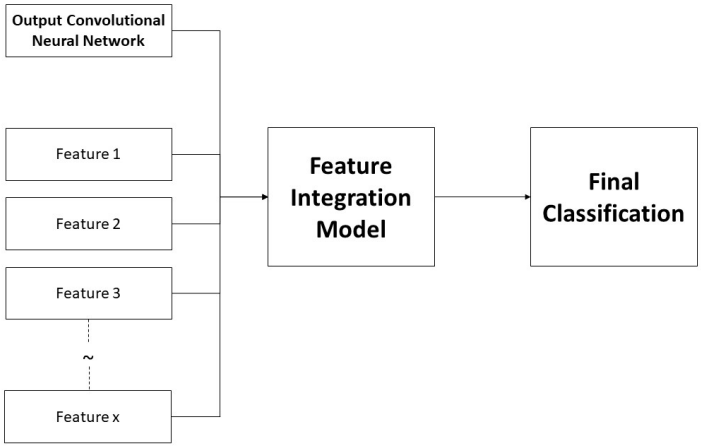

Figure 0.2: Lay-out of the integration of additional features

Results & Discussion

The convolutional neural network is able to make a prediction between 6 categories at an average accuracy 64.6%. Especially on AK and Naevi, the sensitivity is really high. However, the prediction comes with several flaws. Firstly, an imbalance exists between the different categories, since Verrucae & Inflammatory Dermatoses are predicted quite badly. Secondly, it appears that the CNN is overfitted on the categories that indicate skin cancer. The CNN is able to detect skin cancer with a sensibility of 92.3%, but a specificity of only 70% which is a big difference. When the binary CNN is trained that only makes a distinction between skin cancer and no skin cancer, this imbalance is already reduced. Overall, the multi-class CNN is able to make a good initial classification that can be improved by the inclusion of other techniques.

A single decision tree is not a good technique to improve the initial classification from the neural network. Nevertheless, still an overall accuracy of 62% is achieved, but this is mainly achieved by the initial prediction scores that account for 87% of the decisions that are taken. However, the random forest shows ability to improve the results of the original CNN, although it is not much with an average accuracy of 67.5%. The results of the random forest are more balanced as the recall of AK has reduced slightly and the recall of most of the other categories has improved.

Table 0.1: Comparison of different multi-class classifiers

Sensitivity	Verrucae	ID	Naevi	AK	BCC	SCC	Accuracy
CNN	43.2%	**57.9%**	76.9%	**85.4%**	64.4%	54.7%	64.6%
Decision Tree	50%	50%	77%	84%	56%	55%	62%
Random Forest	**56.8%**	55.3%	**79.5%**	80%	**66.1%**	**64.3%**	**67.5%**

In order to classify non-melanoma skin cancer in general, binary logistic regression and binary random forests are the most suitable options. Both methods significantly increase the average accuracy of the CNN and are also less overfitted on skin cancer, especially the binary random forest. The highest sensibility is reached by the logistic regression, whereas the binary random forest scores highest on specificity. Table 0.2 clearly displays this.

Table 0.2: Comparison of different binary classifiers for the detection of non-melanoma skin cancer

	Sensitivity	Specificity	Accuracy
Binary CNN	89.1%	76.0%	83.4%
Multi-class CNN	92.3%	70.2%	82.6%
Binary RF	87.8%	**86.8%**	**87.4%**
Multi-class RF	89.7%	78.8%	84.8%
Logistic Regression	**92.9%**	80.1%	**87.4%**

Finally, logistic regression can also be used to detect AK, BCC & SCC individually out of all

the different categories. AK reaches the highest sensibility at 92.7%, but the specificity of 78.7% is relatively low. SCC shows the opposite, with a specificity of 95.3%. However, the sensitivity is only 73.8%. This result can be improved towards a more balanced result by adjusting the threshold of when the lesion is considered skin cancer. For BCC, this correct threshold seems to be achieved, which is reflected in a better balance between sensitivity and specificity: 83.1% sensitivity at 86.7% specificity.

Conclusions

The research has shown some promising results, especially for the distinction between skin cancer and no skin cancer. The CNN has proven to be capable to make a decent initial classification in the distinction between the 6 categories. Afterwards, both random forest and logistic regression have proven to be capable to integrate the additional features, which has resulted in an improved result of this initial classification. Some of these features have shown more impact than others. The feature "skin type" has shown no significant impact, whereas "color" and "location" have shown significance in various models.

In combination with the individual binary logistic regression results for AK, BCC & SCC, the binary random forest or the logistic regression can be a suitable approach to detect non-melanoma skin cancer. First, a prediction of the probability of having non-melanoma skin cancer can be given, after which a specific probability for AK, BCC & SCC can be displayed. This might result in an application that can be used by both doctors and patients.

Nevertheless, some limitations exist that open the doors to future research. The addition of extra features that have proven to have an impact in literature as well as the use of additional techniques to integrate heterogeneous data will probably lead to more accurate results. Moreover, the collection of extra data will also help to improve the performance of the different models.

Preface

As this research project approaches the end, so does my life as a student. Having studied here in Eindhoven at the TU/e has been an amazing experience and without the support of some very important people, it would not have been the same. Not only have I experienced this full support during the course of this thesis project, but also during all these other years at the university. Therefore, I want to dedicate this short section to thank those people.

First of all, I'd like to thank my direct supervisors that have been of tremendous help during the course of this project: Anna Wilbik, Gertruud Krekels & José van der Waa. Anna, it has been a pleasure to do my thesis under your supervision. You're experience has been of great help and I have experienced a lot of support from you from the very beginning. If I got stuck, our discussions always let to useful news insights that made the process go very smoothly. The fact that we were on a first name basis from the beginning lowered the bar for me to bother you with questions. Thanks for the support and guidance throughout the project from the beginning to the end. Gertruud, you were the one that came with the problem that let to the development of the project. You are aware of the need and the potential of machine learning within dermatology and that has let to a continuous support during the entire process. I admire your energy and ambition and I want to thank you for the opportunity at MohsA. José, you have been an incredible support for me during the project. As my day-to-day supervisor, you were always the one I could easily contact if I had questions. You have done a very admirable job during the data collection and you have taught me everything about the medical aspects in this project. Thanks for everything.

With regard to my time at the university I have to say thanks to my friends who have always been there for me during all those years. Not only have we worked very well together over the course of many course project, my time outside university would not have been the same without you. I know I have made friends for life here and I will be thankful for that forever.

Finally, I have to say thanks to my family. Sanne & Maarten, thanks for being there every time I needed you to. Mom, dad, I know you love and support me in everything I do in every possible way. If there is any issue, I know I can always come to you and you have been supportive during my entire studies. Words can not describe how grateful I am that you're my parents. Thank you!

Frank Hoepel

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem context

Skin cancer is the most upcoming type of cancer in modern societies, especially in countries where a large portion of the population consists of fair-skinned people [Furdova et al. (2020)]. Several types of skin cancer exist and have different levels of severity. A distinction can be made here between melanoma and non-melanoma skin cancer (NMSC). Melanoma is a malignant and fast growing form of skin cancer developing from melanocytes. Although only 2% of the skin cancers is a melanoma, it accounts for most of the deaths [Linares et al. (2015)]. For that reason, a lot of the research regarding skin cancer has been focused on melanoma, even tough NMSC are more common.

The most frequently occurring types of NMSC are basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and actinic keratosis (AK), although the latter is rather considered as an early stage of SCC [Vimercati et al. (2020)]. BCC is the most frequently diagnosed type of skin cancer [Marzuka and Book (2015)]. It is mostly caused by excessive sun exposure and fair-skinned people are more at risk [Lomas et al. (2012)]. The cancer develops mostly at older age, especially after the age of 60 [Furdova et al. (2020), Brash et al. (1996)]. Incidence rates of the BCC are increasing worldwide. For example, in the Netherlands one in five people is estimated to get BCC somewhere in their life and that number appears to be only increasing [Flohil et al. (2011)] These numbers are even greater in countries like the UK and Australia [Lomas et al. (2012)].
SCC is the second most frequent type of NMSC, with an estimated one out ten Americans being diagnosed once in their lifetime. [Kallini et al. (2015)] As explained, the lesion mainly develops from AK. SCC can also be caused by other means, such as an open wound or inflammation.[Madan et al. (2010)] SCC are most often painful, bleeding lesions that appear within weeks. SCC can metastasise in about 5% of the cases and therefore suspicion of SCC requires consultation on short term [Kallini et al. (2015)].
AKs are spots on the skin caused by excessive sun exposure, with older and fair-skinned people being more at risk. The study of Flohil et al. (2013) showed that 37,5% of people aged 45 or older suffers from AK. Since AK does not often develop into squamous cell carcinoma, these spots do not need to be removed necessarily. However, since the progress of AK can't be predicted, the lesions should be regularly monitored [Fernandez Figueras (2017)].

Although NMSC are less severe and have lower mortality rates compared to the aggressive melanoma, a lot of the capacity at hospitals and dermatologists is spent on NMSC. The treatment of NMSC patient is very costly. Between 2007-2011, the annual costs of NMSC treatment in the USA were $8.1 billion each year [Guy Jr et al. (2015)]. These costs consist of both consultation costs and costs of treatment. Patients often demand immediate consultation about the spots on their skin, because they are afraid it could be immediately life-threatening. These consultations are executed on location and cost a lot of time. This leads to planning problems and a waste of

potentially useful time.

Therefore, solutions are urgently required for this problem. A potential solution would be methods that patients would be able to check for themselves whether they have skin cancer or not. This could mean that a patient makes a picture and enters some basic information about common risk factors for skin cancer, such as skin type or age, which would lead to a basic consult or risk estimation for the specific patient. Therefore, a desire exists for a model capturing this process of combining picture input and additional numerical information. Successful implementation of such a model would lead to significant labor reduction for hospitals and dermatological clinics and enables better long-term planning. However, multiple questions arise when approaching this problem. First of all, can a model be developed to successfully classify images of NMSC? How accurate can the results become? Besides, what features are the most significant for the different types of NMSC? And how can those features be extracted successfully? And perhaps the most important question is how these features can be integrated successfully with the results of image classification.

Hence, the reasons to write this thesis. The goal of the thesis project is to build a model that combines the outputs of an image classification model with common features of NMSC in order to successfully classify the images of NMSC. In order to achieve this goal, the thesis has been split in two major parts that are heavily integrated. First of all, a first classification of different types of NMSC will be made based on the pixel values of the different images. A convolutional neural network will be built to achieve this goals, which will be described in detail in terms of design, parameters and results. Afterwards, several methods will be evaluated that intend to optimize the initial classification made by the neural network. These methods will be compared and out of the different methods the most suitable ones will be selected. This should lead to a final recommendation on what is the best method to classify NMSC based both on image and patient characteristics.

1.2 State of the art

In order to retrieve a correct insight in the existing techniques to achieve image classification and the integration of these results with numerical features, two main questions are formulated to retrieve more insight insight on this issue. Firstly, which image classification techniques exist for the classification of medical images. Secondly, what techniques exist that combine image classification with numerical data. The main findings of this literature study will be used for the formulation of the main research questions and the design of the methodology.

In order to retrieve an elaborate answer to the first main question, several methods have been explored such as support vector machines (SVM), random forests (RF) & artificial neural networks (ANN). All three of these methods have proven significant results in the field of image classification [Cracknell and Reading (2014), Khatami et al. (2016), Maxwell et al. (2018)]. However, convolutional neural networks (CNN) have proven the most promising results lately, especially in the field of medical imaging and remote sensing. Since Krizhevsky et al. (2012) built the AlexNet

model, which is a groundbreaking model in the research area, a lot of research has been conducted on CNNs for medical image classification. For example, Brinker et al. (2019b) have developed a model that outperforms most experts for the classification of melanoma skin cancer. Similar results have been achieved in the area of NMSC. Marka et al. (2019) compare thirty-nine studies on NMSC, which show high accuracy results on average. However little research is done on how the results of a CNN can be improved by numerical feature data.

The answer to the second main question is retrieved by focusing on the combination of image classified data and numerical data. The literature research shows that several techniques exist for the combination of data of different modalities and the integration of outputs of different models. A distinction between two types of techniques can be made here. First of all, multi-modal deep learning & information fusion are sophisticated techniques that have shown significant results lately. Especially multi-modal deep learning is very well applicable to integrate multiple variable type [Zhang et al. (2018)]. Since this technique is mostly used together with CNNs, a fine integration exists with the main results from the results of the first part of this research. Secondly, various techniques exist to combine the outcome of different models. In order to use these techniques, the outputs of the different models do not have to be of a different modality necessarily. These methods are called ensemble methods and have proven to be very flexible and improve accuracy [Rokach (2010)]. A very suitable and easily implementable example of an ensemble method is majority voting, where the final output is the most selected output out multiple different models [Yang et al. (2016)]. Besides, the described random forest and support vector machines are applicable for these type of problems as well [Montantes (2020), Maxwell et al. (2018)]. Finally, ordinal logistic regression has proven valuable results in various medical areas [Tietjen et al. (2007), Wu et al. (2015)].

1.3 Stakeholders

1.3.1 MohsA

MohsA is a dermatology clinic located in Eindhoven and Venray, the Netherlands, specialized in the Mohs micro-graphic surgery (MMS), which is a type of surgery used for treating common types of skin cancer. Currently, MohsA suffers from scheduling problems for their clients. Although most of the time the spots the clinic encounters are not skin cancer, and even if they are, do not require immediate surgery, these patients still require appointments. In order to reduce this, MohsA wants to be able to give better consult from distance. An application that already exists is called the OddSpot. This application consists of a form of 14 questions related to a spot on the skin. The result of this form is a probability for the spot being actinic keratosis or basal cell carcinoma.
However, patients using this app often don't feel this is enough. For example, if a score of 3% rolls out, these patients still don't feel safe and still apply for physical consult. Therefore the desire has been formulated to extend the possibilities for distance consult. An opportunity for this would be the ability to provide valuable consultation based on a picture that a patient takes of the concerning lesion. If this can be done successfully by an image classification model, MohsA

can be left out partially, which will reduce the workload of the employees of MohsA significantly. Besides, if this can be combined with already existing numerical models, the consult that is given from a distance can even be improved.

1.3.2 Eindhoven University of Technology

The university is the other big stakeholder in this project. The most important aspect of this project for the university is the scientific relevance of the project. This scientific part is the exploration of the opportunities for integration of heterogeneous data. As explained in section 1.2, several methods exist for likewise problems, but no literature has been found that shows similarities with the way this specific problem is approached. Therefore, if the project is executed successfully, a unique contribution to the existing research can be made. Perhaps these results can be implemented and used in other contexts as well, if this research provides promising results.

1.4 Research Questions

In this section, the research question and its sub-questions are being discussed. These questions are based on the problem as described in section 1.1, the current state of the art as described in section 1.2 and the demands of the different stakeholders as described in section 1.3. Therefore, the main research question is formulated as follows:

How to classify non-melanoma skin cancer based on image and patient characteristics?

The main question captures the overall goal of the thesis project, which is to develop a model in which images are first classified and the results are improved afterwards using heterogeneous data. In order to provide structure in the research process, three sub-questions have been formulated as well. Following these structural steps correctly should lead to a well-considered answer to the main research question.

1. What is the quality of image classification for non melanoma skin cancer?

The first sub-question relates to the classification of images of non-melanoma skin cancer. The goal is to develop a model that gives an initial good value based on the pixel-values of the images. Although multiple techniques exist to do this, convolutional neural networks will be used for this. As explained in section 1.2, convolutional neural networks are the most promising method to do this. The experimental results are used to provide an answer to this sub-question.

2. How to combine image classified data with numerical data?

This sub-question relates to the next step in the process, which is the integration of the numerical data with the initial classification that has been made with the convolutional neural network. In order to give a sufficient answer to this question, several methods will be tested that are able to integrate numerical data with the initial classification from the images. The numerical data comes from the extracted patient characteristics. A selection of the methods that will be tested has been

made in advance and is mostly based on literature. The main criteria in this method selection were whether it is feasible to implement & its effectiveness. The answer to this sub-question will be based on experimental results.

3. Are the models good enough to be used in practice?

The final sub-question relates to the conclusive part of this research. A correct reflection on the results is required to provide a good answer on this question. Therefore, it is important to use the correct evaluation methods for both the image classification model as well as the different models that integrate the numerical data.

1.5 CRISP-DM Methodology

The CRISP-DM methodology will be used as a guideline throughout the research project. The CRISP-DM methodology is a methodology that aims to provide structure in a data-oriented process. [Wirth and Hipp (2000)] Since both parts of the research are designed around data-analysis, CRISP-DM methodology is very well applicable in this project. The design of the CRISP-DM methodology is displayed in figure 1.1.

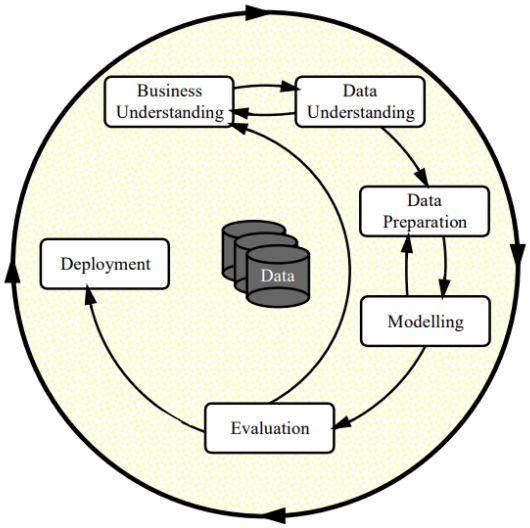

Figure 1.1: The CRISP-DM methodology as described by Wirth and Hipp (2000)

The CRISP-DM methodology will be followed twice. Firstly, in order to answer the first sub-question, and afterwards to answer the second research question. The reason to execute the CRISP-DM circle twice is because for the biggest part the two parts of the research project are very different. Different data is used, different evaluation methods are applied and the goals of both parts differ. Only the business understanding part is the same for both research parts, because the company is the same and the eventual research goal is also the same for both parts of the research. An elaborate overview of the research methodology is provided in chapter 3.

1.6 Research Goals

The goals of this research project are twofold, since the research has a both academic and practical relevance. The academic goals refer to the scientific relevance of this study and its contribution to the already existing literature by filling the scientific gap of methodologies that can classify NMSC using heterogeneous data, which has been described in section 1.2 and chapter 2. The practical goals refer to the added value for MohsA in their trajectory towards reducing consultation time. Below, an overview of the most important research goals have been formulated for both categories:

Academic
- Provide insight in good models for the classification of NMSC.
- Provide insight in the quality of models that integrate features of NMSC with image classified data.
- Provide insight in which features impact the classification of NMSC the most.

Practical
- Provide a model where the input from an image together with the input of a few questions can accurately classify different types of NMSC.
- Provide insight in which features impact the classification of NMSC the most.
- Provide insight in the feasibility of the development of a working classification application.
- Minimize total number of clients by removing unnecessary consultations in order to reduce total waiting time.

1.7 Scope of the project

The scope of a research project is defined by a combination of several factors. The balance between company needs, available time, available resources & expectations should be determined.
This project is bounded by the time for the execution of a master thesis project, which has been determined by the university. Therefore, this project is bounded in a time scope of +/- 6 months. Secondly, the desire of MohsA is to eventually have a model that is able to combine image classified data with additional features, which has been explained in more detail in section 1.3.1. In order to make sure this can be achieved within the limited amount of time, only features that can be extracted from the images have been incorporated in this research, since this bounds the data collection phase of the project to an acceptable time. The time for feature extraction is bounded by the available time of the people at MohsA to help making a useful judgement of the extracted features in the pictures.
The described images of the lesions are data of two sorts. Firstly and primarily, images from the patients have been retrieved from the database at MohsA. Patient privacy is very important at MohsA, so irrelevant patient details have not been shared with the author. Aside from that, images from the internet have been retrieved following strict guidelines to do this safely. The time for data collection is bounded within the available time of the employees at MohsA to work through the database.

2 Background Information

This section is dedicated to provide an in-depth overview of the different types of models that have been used in this research project and literature background in the research area. Therefore, a general introduction to artificial neural networks is described. Artificial neural networks are a machine learning technique that can be regarded as the predecessor of convolutional neural networks, which is currently the most upcoming and most efficient machine learning technique for the classification of images. Moreover, convolutional neural networks are the technique used in this project for the initial classification of images. Therefore, an introduction to the basic concepts of artificial neural networks is required, before the fundamentals of the convolutional neural network can be explained. The different aspects of convolutional neural networks is explained in detail and an overview of the use of convolutional neural networks in the classification of medical images will be provided. Secondly, the different methods are explained that are used in the second part of the project, which are meant to boost the initial classification of the CNN by making use of additional features.

2.1 Artificial neural networks (ANN)

Artificial neural networks are a machine learning technique that mimic the way neurons in the human brain function [Amato et al. (2013)]. A human brain consists of millions of neurons that can be activated by environmental input. For each neuron a certain threshold exists on whether it will 'fire' or not. All these neurons form a complex structure of multiple layers consisting of millions of neurons that either fire or not. The totality of fired and non-fired neurons makes up the way the way humans interact with the perceived environmental information [Graupe (2013)]. This highly complex process is simplified by so-called artificial neural networks (ANN). An ANN consists of an input layer, followed by multiple hidden layers and an output layer. These layers consist of multiple nodes, the term used in ANNs for neurons. Each node in a layer is connected with each node in the previous and following layer, which creates a highly complex system of layers and nodes. The amount of layers and the amount of nodes per layer define the depth of a neural network. More layers and more nodes per layer mean more different parameters and therefore a deeper network. The amount of nodes in the output layer is usually much smaller, because the amount of nodes in that layer translate to the amount of output values. For example, a regression analysis requires one output node, a binary classification between cats and dogs requires two nodes and an analysis with 5 possible outcomes requires 5 nodes in the output layer.

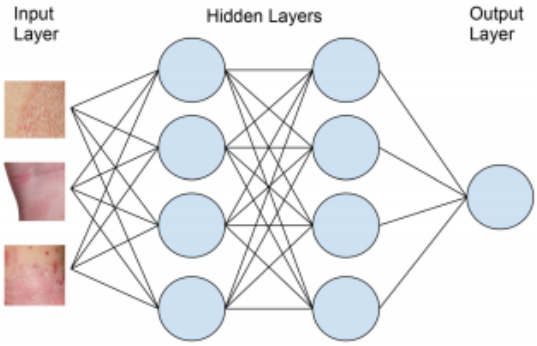

Figure 2.1: Basic structure of an artificial neural network for the classification of skin lesions. [Hogarty et al. (2020)]

Figure 2.1 shows an example architecture for the classification of skin lesions. Information enters the network via the input layer and is fed through the network. The nodes have weights that determine a prediction of the final classification together, after which is evaluated how far away this prediction is from reality. Through a mathematical process called back-propagation is decided how far away that prediction is from the actual value. This process determines how the weights of the nodes in the network are altered.. After multiple runs, the network is trained well enough to make a logical distinction between different classes.

2.2 Convolutional neural networks (CNN)

The CNN is a variant or extension of the classical ANN and is considered a class of feed-forward neural networks, which is mostly being applied in image classification. CNNs consist of multiple layer types: fully connected layers, the main structural component of the classical ANN, pooling layers & convolutional layers. These different types of layers reduce the size of the images and therefore training time is much more shorter than for a classical ANN. Current research on CNN shows the significant improvement that is being made in this field. Previously, random forests and support vector machines have proven to give results at least as accurate as CNNs for image classification, such as the article written by Cracknell and Reading (2014) that showed that random forests and support vectors machines outperformed convolutional neural networks in remote sensing. However, nowadays CNNs have become the most described deep learning tool for image classification. There are several differences between the classical ANNs and CNNs. Most importantly, the layers in the ANN consist only of fully connected layers, whereas the CNNs also consist of convolutional layers and pooling layers. Since these other layers have a significant impact on the reduction of the computation time, CNNs generally have a much shorter computation time than ANNs.
Apart from the reduced computational time, several other advantages of a CNN exist as well.

Firstly, the CNN allows for exponentially increasing capability [Zhao et al. (2019)]. This allows for the development of very advanced models such as AlexNet [Krizhevsky et al. (2012)] and VGG16 [Simonyan and Zisserman (2014)], which have been highly referenced as very well performing and advanced neural networks. Secondly, the hierarchical feature representation can be learned automatically from the data through all the different, multi-level hidden layers. This simplifies the task for the designer drastically. However consequently, the CNN mainly remains a black box, which is probably the biggest disadvantage of CNNs. [Zhao et al. (2019)]

2.2.1 Convolutional layers

In conventional ANNs, each node is connected to all the preceding and following layers. However, for complex tasks like image classification that require a lot of layers for correct classification, these ANNs will become too big computationally. Therefore, in convolutional layers, not every node corresponds with every previous node, but only with a small subset. This is instead of the fully connected layers that are being used in artificial neural networks that preceded the convolutional neural networks. Convolutional layers reduce computational power and the amount of training data required. When a convolutional layer is added, the size of the image shrinks because of the few input values they consume. A filter moves over the spatial dimensions of the layer, that 'convolves' the values of the previous layer using the dot product of the different vectors. For computational reasons, the layer is zero-padded, which means a boundary of zeros is added to do the convolutions successfully.

2.2.2 Pooling layers

Convolutional Neural Networks also have pooling layers, which are meant to reduce the size of the image. The main purpose of using pooling layers is to reduce computation time, but also to avoid over-training of the network. A common method to do this is max pooling. In max pooling the pixel with the highest value of the input value is taken from a subset of pixels in the image. The other pixel values are dropped and only the pixel with maximum value maintains. Alternatives for max pooling exist. An example is average pooling, where the average value of the selected pixels is taken. However, this requires more effort computationally, for which reason max pooling is the more commonly applied approach.

2.2.3 Fully connected layers

In the end, fully connected layers are also part of a CNN. However, the amount of fully connected layers is significantly lower than in CNNs, which is the main distinction between the two methods. In widely known networks such as AlexNet, which was one of the first well-functioning CNNs, the convolutional layers are often alternated by fully connected layers. In total, AlexNet consists of three fully connected layers and five convolutional layers, which are sometimes followed by max pooling layers [Krizhevsky et al. (2012)]. The full design is presented in figure 2.2

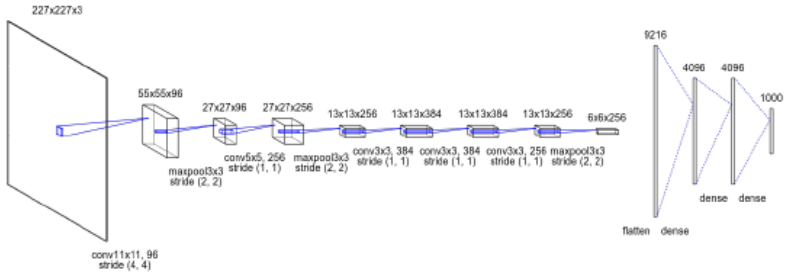

Figure 2.2: Architecture of AlexNet [Krizhevsky et al. (2012)]

2.3 Convolutional neural networks for medical image classification

The different papers each have a different area in which the research has been applied, although some of the researches overlap. For example, Brinker et al. (2019b), Brinker et al. (2019a) & Haenssle et al. (2018) all approach the issue of detecting skin cancer using a CNN. Contrary to the non-melanoma skin lesions context of this research, the mentioned papers report on the detection of melanoma skin cancer. In Brinker et al. (2019b) researchers trained a CNN using a pretrained ResNet network. Afterwards, the researchers showed that the network performed better than 87% the 157 dermatologists at detecting the melanoma from the test. The authors achieved similar results when this network was exposed to a clinical image set without having been trained on medical images in Brinker et al. (2019a). The CNN showed a lower variance in the results compared to the dermatologists that classified the training set, which indicates a high robustness of the model. Previous research also showed valuable results. The model described by Haenssle et al. (2018) outperformed a majority of the dermatologists that classified the images of the test set. The authors made use of a pretrained GoogleNet Inception CNN, which had additionally been trained with a total training set of 100,000 images.

The use of CNN in other medical areas has been explored as well. Hemanth et al. (2018) built a relatively simple Deep CNN with consisting of one convolutional layer, one max-pool layer and one fully connected layer. This network has achieved an accuracy of 94.5%. Afterwards, this networks as been modified by adjusting parameters in the fully connected layer and in the convolutional layer which lead to an accuracy increase to 96.4%. In Bejnordi et al. (2017) several deep learning algorithms that detect lymph node metastases have been compared. These algorithms have been developed for a competition between different deep learning techniques called CAMELYON16. A total of 32 algorithms have been compared in the paper of which 25 were deep learning models. Most of those deep learning models are CNNs. These CNNs generally outperformed the other methods.

2.4 Additional Techniques for Feature Integration

After the CNN has been developed, several characteristics of the lesions are retrieved from the pictures. Those can only be features that can actually be retrieved from the images themselves.

Examples of such features are position on the body, elevation and shininess; features that are mentioned by Van Der Geer et al. (2015) as indicators of several NMSC. Thereafter, a few different methods will be described that combine the prediction scores of the classification model with the presence of different features. These techniques are capable of handling these different types of information and appear to be the most suitable for this research project.

2.4.1 Decision Tree

A decision tree is a machine learning method that is able to make a classification based on exclusive decisions for the independent variables. Following a path through these decisions leads to a final value for the dependent variables. A clear benefit of a decision tree is its clear visualisation, as it can be presented as an upside-down tree with leaves. Figure 2.3 shows an example decision tree for the determination of the mode of transportation depending on the weather.

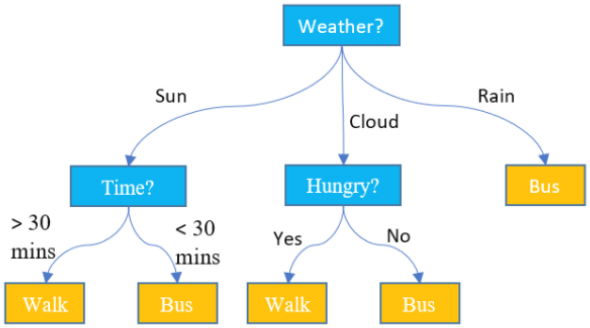

Figure 2.3: Example of a simple decision tree

For this project, the decision tree is selected because it is able to incorporate the prediction values of the convolutional neural network as well as the values of the different extracted values. By this means, the decision tree should be able to improve the original result of the convolutional neural network by inclusion of the different features. The decision tree is not expected to provide very big improvement, but will mostly be used for a comparison with the random forest.

2.4.2 Random Forest

A random forest is an aggregation of multiple decision trees. These trees all follow different paths towards a final classification based on the independent variables, after which these classifications are all aggregated. The final classification is based on the majority vote of the individual classifications of the decision trees. The main purpose to use a random forest over a single decision tree is to overcome the weakness of uncertainty of one single tree. Therefore, random forest is more robust and more reliable than a single decision tree. The main reason a single decision tree is included in this research is because of its great visualization possibilities.
In the context of this project, the random forest should be able to improve the original result

from the convolutional neural network. The random forest is able to incorporate both the results from the convolutional neural network and the values of the extracted features with the intention to retrieve accurate values. The expectation is that the random forest will outperform the single decision tree.

2.4.3 Logistic Regression

The final techniques that will be implemented is logistic regression. Logistic regression will be used for the determination of relationships between variables in binary classification methods. Although the core of the project focuses on multi-class methods, also binary models will be compared in order to make a distinction between skin cancer and no skin cancer. Moreover, this techniques will be used to detect the specific types of skin cancer out of all the other categories.

Logistic regression differs from linear regression for a few reasons. Firstly, in linear regression the output value is continuous and can therefore take every value on the range of infinite possible results in theory. Logistic regression is bounded to values between 0 and 1 and can only take values within those range. Secondly, linear regression is mainly used when the dependent variable is continuous, whereas logistic regression is used when the dependent variable is categorical. In this research, the dependent variable is whether the lesion represents skin cancer or not. Therefore, logistic regression is the most suitable regression methods for this research.

2.5 Evaluation Methods

Various performance measures exist to evaluate the performance of a model. The following evaluation methods have been selected to measure the performance of the different models, because they are considered most relevant for this research.

2.5.1 Confusion Matrix

The Confusion matrix is a graphical representation on how a model classified the cases. The x-axis shows how the different cases have been classified. An example of a confusion matrix in a binary classification model can be found in table 2.1

Table 2.1: Example of a confusion matrix with two possible outcomes

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The lesions can generally be classified in 4 different categories: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). True positive values are cases that have

been correctly classified as being that case. True negative values are cases that have been correctly classified as not being that case. False positive are cases that have been wrongly classified as being that case, while it is actually not. False negatives are cases that have been wrongly classified as not being that case, while it actually is. From this point the abbreviations for the different categories will be used throughout this report, both in text and in formulas.

2.5.2 Accuracy

The first evaluation method is accuracy, which is a very generally applied evaluation method. It is defined as the ratio of correctly classified cases amongst all cases, both true positive values and true negative values. Mathematically, the formula looks as follows for a binary classification problem:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

In a multi-class environment, accuracy is considered as the average accuracy per class. In formula:

$$\frac{\sum_{i=1}^{k} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + fN_i}}{k} \tag{2.2}$$

A general advantage of accuracy is that it is an easily interpretative measure that generally gives an indication of the overall performance of the network. Therefore, in this project it is used to make a general shift between good and bad models. However, a major disadvantage is that no distinction is made between the classes and generally a flawed result in an imbalanced data set is given. For example, if one of the classes consists of 50% of the data, an accuracy of 50% can already be retrieved by simply just predicting that specific class. Therefore, other evaluation methods are required to get a more accurate indication of the performance of a model.

2.5.3 Recall & Precision

Recall is the ratio of correctly predicted values per class. Mathematically, it translates to the amount of true positive values (TP) amongst the sum of true positive values and false negatives (FN), which is the sum of the real positive values of a class. In formula:

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

Precision is the term for the ratio of correctly predicted positive values out of all positively predicted values of a class. Mathematically, it translates to the amount of true positive values (TP) amongst the sum of true positive values and false positives (FN), which is the sum of all the predicted positive values. In formula:

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

Compared to accuracy, these values give a more detailed indication of the performance of a model as it considers each class separately. Moreover, it can give information on whether a model is

overfitted or underfitted on a certain class. Two examples can illustrate this. Firstly, if average accuracy is high, but the recall & precision values show that one or two classes account for most of that accuracy, while other classes are behaving poorly, the model is not as good as one would expect from the accuracy value. Secondly, if one class has a very high recall, but a very low precision, the model is probably overfitted on that class. It predicts that particular value very often, to make sure no samples in that class are missed. However, this causes other classes to perform poorly. Therefore, recall & precision ideally should be balanced.

2.5.4 F1 score

The F1 score is a commonly used evaluation method to convey a good balance between precision and recall. The mathematical formula is as follows:

$$F1score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{2.5}$$

Although the F1 score can be seen as a balance between recall & precision, it is not an average value. A big imbalance between the scores for recall & precision will lead to a low F1 score. A situation with a similar average value of recall & precision, but with these values being more balanced, will lead to a higher F1 score. Conclusively, the F1 score is a good measurement for big discrepancies between recall & precision

2.5.5 Sensitivity & Specificity

Since this project is embedded in a medical context, sensitivity and specificity will be used as the most important evaluation measures. Sensitivity is actually the same as recall. Therefore, sensitivity and recall will be used interchangeably.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.6}$$

Specificity translates to the fraction of correctly predicted negative values, true negatives (TN) amongst all predicted negative values. This translates to the following formula:

$$Specificity = \frac{TN}{TN + FP} \tag{2.7}$$

The main purpose of using specificity is to make sure that people who don't have a condition are qualified as healthy. In this case, specificity has use in the multi-class research, but has most purpose in the binary research where skin cancer is compared with no skin cancer.

2.5.6 Gini Impurity

The Gini impurity is an evaluation method that is mostly used for the evaluation of decision trees. Named after the Italian scientist Gini, it gives an indication on how likely a random sample that

enters the model will be incorrectly classified. The formula to compute the Gini-coefficient is as follows:

$$G(k) = \sum_{i=1}^{J} p(i) * (1 - p(i)) \tag{2.8}$$

with $p(i)$ being the probability that item i out of set k is classified correctly. The Gini impurity reaches its optimum at 0, when all items are being classified in the same category. The Gini impurity will not be used a lot in this thesis, only for the evaluation of the robustness of the single decision tree.

3 Methodology

This section is dedicated to describe the methodology of the research project in detail. Firstly, the general research lay-out is described. This is an important step, because the research is built up in two parts that are highly integrated. The division in two parts aims to provide a clear structure, since each of those parts aim to answer one of the first two research sub-questions. After the research lay-out description, a general data description is presented. In this section the different output classes, the data collection & the data preparation are described in detail. Moreover, an elaborate description of the extracted features will be provided in this section as well. The reasoning on the decisions for the use of the specific output classes will be treated in this section as well. Therefore, this step is related to both the data understanding and data preparation steps in the CRISP-DM cycle. The final two sections are dedicated to the description of the methodology for the development of the two models. The first model captures a convolutional neural network that is able to make preliminary classification of the different images. The process what the relevant parameters are and what will be altered during the modeling process is described in detail, as well as the relevant hyperparameters that need understanding before starting the modeling process. Moreover, the relevant evaluation methods are described here as well. In the second model, different methods will be tested to boost this initial classification from the CNN by integrating additional features. To execute this part of the project successfully, a general lay-out is required to maintain consistent modeling. Similar to the first model, an overview of the relevant evaluation methods has been described here as well.

3.1 Research Lay-Out

The research consists of two parts. Firstly, an initial classification will be made of the images in the database. This database consists of pictures of different types of non-melanoma skin cancer and pictures of other lesion. The characteristics of the database will be treated in section 3.2. Secondly, this initial classification will be improved with features that have been retrieved from the images. Several different suitable methods will be applied and compared. The results will be compared and interpreted in terms of practical and theoretical relevance. The theoretical relevance relates to the extent to which the results are correct and the models can make correct classifications with the different data types. The practical relevance relates to the extent to which the results are useful to integrate into an application that is able to remotely classify NMSC. Conclusively, this leads to the research methodology design in figure 3.1. The lay-out has been designed according to the guidelines of Wieringa (2014). The goal of preparing such a design is to achieve continuous improvement of the artifact during the research process.

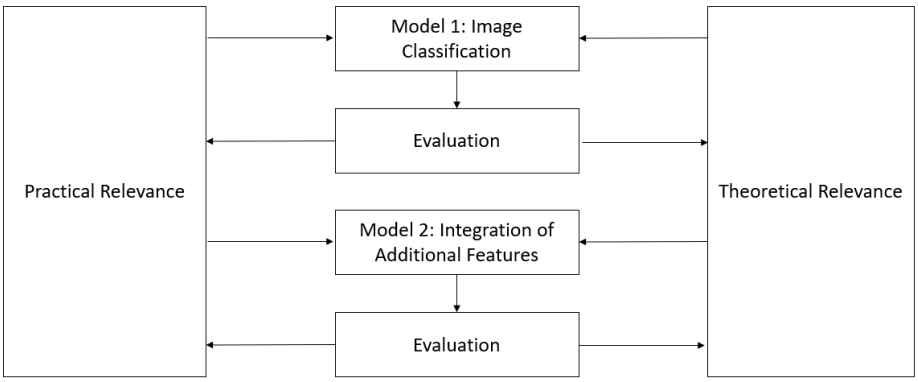

Figure 3.1: Design of the research methodology following the guidelines of Wieringa (2014)

Both parts of the research project will be executed according to the guidelines of the CRISP-DM methodology [Wirth and Hipp (2000)]. As explained in section 1.5, the CRISP-DM cycle will be executed twice. The main reason to approach it that way, is because the two parts of project differ fundamentally. Therefore, most of the steps in the process can not be executed the same way. Nevertheless, some overlap in the process still exists, mostly in the business understanding and the data collection parts of the cycle.

3.2 General data description

3.2.1 Output classes

A distinction has been made between 6 classes. These classes are AK, BCC, SCC, Inflammatory Dermatoses, Naevi & Verrucae. As explained in the introduction in chapter 1, BCC & SCC are NMSC, and AK is a preliminary stage of NMSC. However, for the ease of this project AK will be referred to as either 'skin cancer' or NMSC, despite the fact that this is technically not entirely correct. Those outputs are the most important ones to detect correctly as they form the core of the project. The other categories consist of groups of skin diseases and skin lesions. The first class is Inflammatory Dermatoses (ID) and embraces multiple skin rash types. These are dermatomycosis, eczema, psoriasis, pytyriasis versicolor, pytyriasis rosacaea & urticaria. The class of Naevi consists mostly of regular moles. The last class is the class of Verrucae, which is a class consisting mostly of regular warts. These last three categories have been introduced for the neural network to make a clear distinction between skin cancer and no skin cancer. A neural network can't be trained to make a distinction between non-melanoma skin cancer and 'everything else', because that would embrace an enormous range of different lesions that have no similar patterns at all. Therefore, this group of 'everything else' has been split up in the three subgroups that have shared similarities and define the classes that differ from the non-melanoma skin cancer types. This also means that lesions that can not be fitted in one of the six categories can't be correctly classified by the neural network, since the image can not be assigned to one of the output categories.
For structural reasons, the different categories are presented in the same order in every applied

method. This order is Verrucae, Inflammatory Dermatoses, Naevi, AK, BCC & SCC and is based on the severity of the lesions. Verrucae is considered the least severe and SCC the most severe.

3.2.2 Data collection

The images have been retrieved from two different sources. The major part of the images comes from the internal database of MohsA. The labels of the different images in the database are the official diagnoses of the different experts that work and have worked at MohsA. Therefore, these classifications are very reliable. The other part of the images has been retrieved from different databases of the internet. The classification of those images is less reliable compared to the images of the MohsA database. Therefore, several measures have been taken to increase reliability. First of all, the images have been selected by the dermatologists at MohsA. Only, if they were certain that an image indeed corresponded to the given classification, the image was added to the database. Secondly, the images with a too low pixel value have been removed from that database. A low quality image is no guarantee for a correct classification, so risks should be avoided. Finally, at a later point in time the classifications of these images have been re-assessed by the dermatologists. If this second opinion differed from the initial classification or a lot of doubt arose from the initial classification, the image would ultimately be removed. This entire process has been constantly monitored by weekly meetings with José van der Waa, who has been the connection between the theoretical and operational part of the data collection process. Conclusively, 1764 unique images of the 6 mentioned categories have been collected in total.

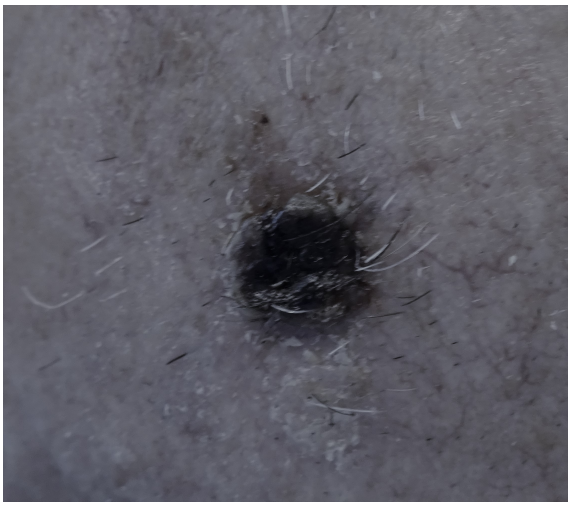

Figure 3.2: Image of a BCC of an anonymous patient from MohsA

3.2.3 Features

The different techniques in the second part of the research make use of the features that have been manually derived from the images. These features include aspects like location on the body, color of the lesion, elevation, etc. These features will be included as input values for the different methods. It is important to get a good overview of these different categories and to make sure

that they are well-represented in the models. Most of these features are categorical, except for the prediction scores, which is a numerical input. Comparable to the process of data collection [section 3.2.2], the manual retrieving of the features has been monitored and verified by drs. van der Waa and dr. Krekels during weekly meetings. However, the reliability of the presence or absence of those features is not 100%. The assumption is that the reliability on to what extent the features are actually present in the image is equal to the reliability of a patient indicating whether the feature is present or not.

3.2.3.1 Location on the body

The different lesions are found all over the body. Since for example AKs, BCCs and SCCs are more significantly present on parts of the body which have been more exposed to the sun, this is a very interesting parameter to consider. [Vandiver et al. (2015)] [Yaldiz (2019)]. For the ease of this project, the amount of different places on the body has been limited. For example, no difference has been made between 'cheek' and 'chin'. This is all integrated in 'face'. Also 'fingers' are integrated in 'hands' and 'toes' in 'feet'. The reason is to limit the amount of classes and make it easier for the different machine learning models to discover meaningful relations between different features. Table 3.1 provides an overview of the different locations on the body that the lesions were found on. Although a lot of different categories exist, not all lesions can be placed in a category. Mostly, this happens because it is unclear from the picture where it is located on the body because the image is zoomed in on the lesion already. Although this is very useful for the image classification model, because the lesion is very clearly visible, the location remains undefined. Therefore, the category 'undefined' has been used as well. During the modeling process, this category will be treated as a separate location like the other categories.

Table 3.1: Frequencies and prediction scores per location from binary CNN

Location	**Nr. of samples**	**Percentage**
Arm	189	10.71%
Back	202	11.45%
Belly	44	2.49%
Buttock	13	0.73%
Chest	58	3.28%
Ear	56	3.17%
Eye	18	1.02%
Face	227	12.87%
Foot	40	2.68%
Hand	156	8.84%
Head	227	12.87%
Leg	83	4.71%
Lips	19	1.08%
Neck	61	3.46%
Nose	82	4.65%
Pubic Area	6	0.34%
Shoulder	29	1.64%
Side	6	0.34%
undefined	236	13.38%
Total	1764	

Most of the lesions are found on the arm, back, face, hand and hands. Other categories are less common. Differences exist in the percentage of skin cancer that is found on the different body parts. For example, 0% of the lesions located on the foot are skin cancer, while 91.5% of the lesions located on the nose are skin cancer. This is an interesting result with regard to the second part of the research in order to investigate which factors influence the prediction of skin cancer the most.

3.2.3.2 Skin type

A distinction between three different skin types has been made: low pigmentation, neutral & high pigmentation. As literature shows, people with less pigmented skin have a much higher chance of being diagnosed with skin cancer [Brenner and Hearing (2008)]. Moreover, white people tend to have a 40% chance to get NMSC or melanoma in their life, which is way lower for Hispanic, Asian or African people [Agbai et al. (2014)]. Therefore, it would be interesting to prove skin type to be a predictor for skin cancer.
The database of images is biased towards people with low pigmentation. The reason for this is twofold. Firstly, the database mostly consists of images taken from patients at MohsA. Therefore, most of the people in the database are ethnically Dutch, which implies a low pigmented skin.

Secondly, as literature shows that people with low pigmentation are especially vulnerable for skin cancer, the people that come to MohsA and let their picture taken are mostly of white color. This increases the bias in the data set. The following table shows the frequency of the different skin types:

Table 3.2: Distribution of skin types

Pigmentation	**Nr. of samples**	**Percentage**
High pigmentation	3	0.2%
Neutral pigmentation	71	4.0%
Low pigmentation	1690	95.8%
Total	1764	

As the table shows, people with low pigmentation are enormously over-represented. Therefore, the expectation is that skin color will not have a huge impact on the improvement of the initial classification when integrated.

3.2.3.3 Color

There is a wide variety of colors lesions can have. For example, Naevi are generally more brown colored, because these lesion consist of over-pigmentation, whereas other lesions almost never show that color. Therefore, the feature 'color' is considered important for this research. Moreover, retrieving the color of a lesion from the images is relatively easy. A wide variety of colors have been retrieved from the data. The following table shows the frequency of the different colors of the lesions:

Table 3.3: Distribution of lesion colors

Color	**Nr. of samples**	**Percentage**
Light Red	914	51.8%
Brown	334	18.9%
Regular Red	283	16.0%
White	137	7.8%
Yellow	76	4.3%
Black	11	0.6%
Dark Red	7	0.4%
Green	1	0.1%
Purple	1	0.1%
Total	1764	

The table shows that a majority of the lesions are light red colored. Moreover, the biggest part of the lesions consist of only five different colors. Black, dark red, green and purple lesions are much less represented.

3.2.3.4 Shininess

Shininess is a binary variable indicating whether the lesion has a shiny appearance. This feature is generally an indication of several types of BCC [Marzuka and Book (2015)].

Table 3.4: Distribution of shininess

Shininess	Nr. of samples	Percentage
Yes	499	28,3%
No	1265	71,7%

3.2.3.5 Scaliness

Scaliness is a binary variable that indicates whether a lesion has a scaly appearance or whether a place on the skin consists of multiple flakes. These flakes are typically white. Scaliness is generally an indication of AK. The scales usually can be found on the skull, face or the lower arm, locations that generally have a large sun exposure [Shoimer et al. (2010)].

Table 3.5: Distribution of scaliness

Scaliness	Nr. of samples	Percentage
Yes	472	26,8%
No	1292	73,2%

3.2.3.6 Red Edge

This feature embraces the question whether the lesion is bounded by a red edge. Just like the previous features, this is a binary variable. Red edges can be found mostly with certain naevi and are an indication of an SCC in combination with the lesion being white on the inside.

Table 3.6: Distribution of the presence of a red edge

Red Edge	Nr. of samples	Percentage
Yes	336	19,0%
No	1428	81,0%

3.2.3.7 White Inside

As said in the previous subsection, a lesion with a white inside is a clear indication of SCC. This is especially the case when the lesion is shiny and has a red edge as well. This feature is also a binary variable.

Table 3.7: Distribution of the presence of a white inside

White Inside	Nr. of samples	Percentage
Yes	415	23,5%
No	1349	76,5%

3.2.3.8 Elevation

An elevation in the skin or in the lesion can be an indication of multiple different skin cancers or skin diseases. This variable is also a binary one. An elevated lesion in combination with the presence of other features can be a good indication of BCC or SCC.

Table 3.8: Distribution of the presence of an elevation

Elevation	Nr. of samples	Percentage
Yes	743	42,1%
No	1021	57,9%

3.2.3.9 Blood Clot

The presence of blood or a blood clot can be an indication of skin cancer. [Lisboa et al. (2016)] Therefore, the presence of a blood clot on the lesion is taken into account as well.

Table 3.9: Distribution of the presence of a blood clot

Blood Clot	Nr. of samples	Percentage
Yes	439	24,9%
No	1325	75,1%

3.2.3.10 Prediction

The final parameter that is included as a skin cancer feature is the prediction score from the CNN. As explained in section 3.1, the CNN makes predictions that are used as input values in the second model. A distinction can be made between two types of input values: Multi-class prediction values and binary prediction values. These values are the results of the two types of CNNs that have been developed, which is explained in detail in section 3.3.

Multi-class prediction values

For each of the 6 lesions, the categorical CNN produces a score between 0 and 1 on the probability per category. All these probabilities add up to 1. This leads to 6 different input values that are added to the other features. For each of the 6 different categories this score represents the probability of the image representing that particular category. Table 3.10 displays an actual example on how such a prediction score is represented.

Table 3.10: Example of predicted probability per category as input values for the models

Predicted probability scores	SCC	BCC	AK	Naevi	ID	Verrucae
Value	0.058	0.034	0.000	0.222	0.000	0.684

Table 3.10 shows that for this particular lesion, the CNN estimates the probability of the lesion to belong to the category Verrucae to be 68.4%. Besides, the model thinks that Naevi is the second most likely category for this images. The other categories are considered less likely and therefore have a very low input score. These prediction probabilities will be produced by the CNN for every single lesion.

Binary prediction values

The binary CNN gives a different output score than the CNN with 6 classes. Since the output layer of the binary CNN only has one node, the prediction score is a value between 0 and 1 on the probability of the lesion being skin cancer or not. Therefore, regarding the methods that will be used to distinguish between skin cancer and no skin cancer, only one column is added as a feature called "Prediction". If the value for Prediction approaches 1, the CNN is very confident that the image concerns a type of skin cancer, whereas if the value approaches 0, the CNN is confident that the images does not display a type of skin cancer.

3.2.4 Data preparation for image classification

The data sets have been divided upfront into three sets: the training set, the validation set and the test set. The test set has been held apart first. It consists of 15% of the unique images and is stratified, meaning that for each class 15% of the images has been put into the test set. The distribution of the classes in the test set is therefore the same as in the total data set. The other 85% is used to build the training set and the validation set. This part of the data preparation part is required for both the image classification model and the feature integration model. However, the image classification model requires additional data preparation. Firstly, almost all the images have been cropped in order to make sure that pictures are clearly focused on the lesions themselves. Prominent parts in the images which are irrelevant for successful image classification are removed so only the lesion is clearly visible. This should make it easier for the neural network to learn about the lesions. If an ear or a nose is too prominently present in the picture, the network might only detect the ear or the nose and ignore the lesion. If this happens to often, the network might detect irrelevant patterns which will lead to inaccurate classifications. Secondly, data augmentation techniques have been applied to increase the total number of images in this data set. The images have been rotated 180 degrees and flipped horizontally. The images in the classes SCC, ID, Naevi & Verrucae have been vertically flipped as well in order to create a balance in the number of images in the different classes. This has not been done for the classes AK & BCC, since these classes consist of more images. Not applying this specific augmentation technique on AK & BCC causes the training set and validation set to be more balanced. After these data augmentation techniques have been applied, table 3.11 has been constructed, providing

on overview of the amount of images in each set. 20% of the images have been put in the validation set, while 80% has been put in the training set. This division is totally random, so contrary to the test set, the validation set is not stratified.

Table 3.11: Distribution of the total data set used for the CNN

	Training set	Validation set	Test set
Verrucae	567	193	44
Naevi	699	192	39
Inflammatory Dermatoses	682	191	38
AK	730	141	55
BCC	801	174	59
SCC	759	169	42
Total	4238	1060	277

3.2.5 Data preparation for feature integration

In essence, the data preparation process for feature integration models has a lot of similarities with the image classification model. Firstly, the test set that has been used to test the quality of the CNN is also applied in the second part of the research, because consistency is essential to produce meaningful results. Secondly, some additional adaptations have been applied in order to balance the data set. However, differences exist in the way these adaptations have been applied. First of all, only the original images have been used and not the augmented ones. The addition of augmented images can help neural networks to learn better, but this is irrelevant in situations where the machine learning techniques handle binary and numerical feature data instead of the image data. A detailed overview of this feature data has been provided in section 3.2.3. Secondly, in order to have balanced input values for each category, copies of images of underrepresented categories have been added to the training set. This means that all categories are equally represented in the training set. This is a major difference with the training set of the image classification model. Finally, no validation has been separated from the training set. This is because validation sets are used to test the results of different models that apply the same technique. This is highly relevant for the CNN, where a lot of different models types will be tested because of all the parameters that can be altered. For the feature integration, not a lot of different model options exist. Therefore, the training set will be kept together to keep enough data to train these models thoroughly.

Table 3.12: Distribution of the total data set used for the feature integration models

	Training set	Test set
Verrucae	331	44
Naevi	331	39
Inflammatory Dermatoses	331	38
AK	331	55
BCC	331	59
SCC	331	42
Total	1986	277

3.3 Model 1: Convolutional neural networks for image classification

The first model that has been built is the image classification model. The building of this model aims to retrieve an answer to the first research sub-question. This section is dedicated to how the model and the parameters are altered in order find the optimal model. The parameters can be separated in structural elements that relate to the development of the architecture of the optimal structure of the CNN and hyperparameters that need to be predefined before training the model. These structural elements are convolutional layers, pooling layers, fully connected layers & drop-out layers and have been explained in section 2.2. Moreover, the amount of nodes per layer will be altered as well. The hyperparameters that will be predefined are the loss function, the optimizer, the activation functions and the image size. Finally, the evaluation methods that have been used to measure the quality of the model will be proposed. This section relates mostly to the modeling and evaluation parts of the CRISP-DM methodology.

3.3.1 Parameters

Convolutional layers

The amount of convolutional layers has a major impact on the performance of the CNN. Too few layers will cause an underfitted network while too many layers can cause an overfitted network. It is hard to prove the optimal amount of convolutional layers, since it is interrelated with the amount of pooling layers, fully connected layers and the amount of nodes per layer. Therefore, the amount of convolutional layers that will be tested are 2 & 3 together with different combinations of the other parameters..

Pooling layers

In order to downgrade the size of the images, pooling layers are added to comprise the images. Because of computational reasons, max pooling is selected as the method to achieve this. The size of the patch that moves over the image to select the groups of pixel has been set on 2 by 2. This means that each time a pooling layer is added, the image will be reduced by a factor 4.
The amount of pooling layers that can be added is not endless. Too many pooling layers will cause the image to become too small to make effective training impossible and the network will become

underfitted. Therefore, the amount of pooling layers that are tested are 1, 2 & 3. Besides, there will not be more pooling layers than convolutional layers.

Fully connected layers

A CNN also consists of fully connected layers, since those can handle more connections than convolutional layers. Not many of these layers are necessary since fully connected layers influence computation time negatively and can cause an overfitted network. Usually, these fully connected layers are placed at the end of the CNN. The number of layers that are tested are: 0, 1 & 2.

Dropout layers

Beside the described layers, another type of layer exists as well, which is the dropout layer. Contrary to the other parameters, the dropout layer is not involved in the classification itself. The dropout layer randomly selects a percentage of the data and removes it from the training process. This can mean that after a dropout layer, the rest of the neural network will be trained using only 80% of the images, as the dropout layer removes 20% of the data. The goal of the dropping out images is to avoid overfitting. Neural networks with 0, 1 & 2 dropout layers have been tested at dropout percentages of 10% and 20%.

Nodes per layer

Finally the amount of nodes per layer is the last structural element that will be altered in order to retrieve the optimal CNN for the classification of NMSC. This element is very important as it influences the amount of parameters in the network drastically. Too many nodes per layer will make the network impossible too train for computational reasons, but also because enough data is required to actually adjust the weights on the different nodes. Too few nodes leads to a network that is unable to train a model to distinguish different parameters.

3.3.2 Hyperparameters

Loss functions: binary cross-entropy & sparse categorical cross-entropy

The loss of a neural network is the parameter that shows how far away the neural network is from the correct classification. For the selection of the loss function two options exist. Firstly, sparse categorical cross-entropy is used if a multi-class classification is required. The losses are calculated per output class. Therefore, this is the required loss function for the multi-class CNN. Secondly, binary cross-entropy is used when a distinction is required between two categories. Here, only one single loss is calculated, since there is only one output class in the neural network that takes on different values for the different categories.

Optimizer: Adam

The optimizer of a neural network makes sure the loss is minimized, which allows the network to be as accurate as possible. As optimizer, the Adaptive Moment Estimation (Adam) optimizer

has been selected, which is the current default optimizing function, because it generally achieves good and fast results. It is an alternative for the classical stochastic gradient descent, which was previously the common optimizing function. In general the Adam Optimizer is regarded to perform superior over other optimizing functions. [Kingma and Ba (2014)] [Ruder (2016)]

Activation functions: ReLU, SoftMax & Sigmoid

In the convolutional neural network, two activation functions have been selected. Between the hidden layers, the Rectified Linear Unit (ReLU) activation function has been selected:

$$f(x) = \max(0, x) \tag{3.1}$$

with x being the input value to a node.
In general, the ReLU is the most commonly applied activation function between hidden layers. The main reason for this, is that the ReLU has the lowest computation time compared to other activation functions such as $f(x) = \tanh(x)$ [Krizhevsky et al. (2012)]. This is because not all neurons are activated at the same time. The function should reach a certain threshold, namely 0, before the function actually starts computing an output value. Therefore, the computation time is less than other activation functions and the best suitable option to efficiently navigate through the different nodes and layers of the convolutional neural network.
For the activation function in the output layer two options exist. Which activation to choose depends on what type of classification is desired. Firstly, the SoftMax activation function will be used for the categorical classification of a multi-class CNN. The formula for the SoftMax function to compute probability x of the image being category i:

$$\text{SoftMax}\,(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{3.2}$$

with e^{x_j} being the prediction score for category j in the sum of all n categories.
The binary classification requires a different activation function. The Sigmoid classifier is the most suitable option for a binary classification and is a deduction of the SoftMax function. The formula of the Sigmoid function to compute probability x of the image being category i:

$$\sigma(x_i) = \frac{1}{1 + e^{x_i}} \tag{3.3}$$

Both the Sigmoid function and the SoftMax function can be used in the output layer of a binary classification model and are technically the same classifier. However, the Sigmoid function is used more often for binary classification because of its simplicity.

Image size

Finally, the image size should be set upfront, which is 64 by 64 by 3. A balance is required between computation time and the amount of information that needs to be put in the CNN. Since

the CNN uses the combinations of pixel values as an input, one might think that a high image size guarantees the best results. However, after a certain image size, the CNN is capable enough to analyze the images. Therefore, extra images would only lead to a higher computation time. 64 by 64 are the amount of pixels horizontally and vertically, while the depth of the image is 3 as we are working with colorized pictures. This color is constructed by a so-called red, green & blue (rgb) layer. The horizontal and vertical length of the image should be equal, since multiple convolutions take place to the image in the network. Since squared images stay squared over these convolutions, the images will stay balanced. The length of 64 has been selected to make sure the images are big enough to make a good prediction, while it is still small enough to maintain a suitable computation time. An image of 100 by 100 by 3 consists of 30,000 parameters, while an image of 64 by 64 by 3 consists of only 12,288 parameters. Therefore, if the image is 100 by 100 by 3, the network will take almost 2.5 times longer to train the model.

3.3.3 Evaluation of the neural network

Several evaluation methods are applied to evaluate the performance of the convolutional neural network. Firstly, a confusion matrix has been constructed of the actual and predicted output classes. This gives a graphical overview of the performance of the model. Afterwards, the overall accuracy is measured in order to get a general idea of the performance of the convolutional neural network. However, this measurement is not accurate enough in itself to establish a correct opinion on the quality of the model. A detailed look into the recall and precision of the different categories should indicate whether the model could be overfitted or underfitted. The combination of these evaluation methods should give enough information to make an educated evaluation about the performance of the model. A detailed explanation on these evaluation methods has been provided in section 2.5.

3.4 Model 2: Integration of additional features

After the image classification model has been developed, the results will be integrated in a model that integrates the additional features. This is the second part of the research as visualised in figure 3.1 and aims to provide an answer to the second research sub-question. As explained, the business understanding and data collection part in this section is the same as for the image classification model. Therefore, this section mostly relates to the modeling and evaluation parts of the CRISP-DM methodology.

3.4.1 Experimental set-up of the modeling process

For the investigation on whether the initial classification of a skin lesion can be improved with numerical and categorical information with regard to the features of the lesions, different suitable techniques are tested. These features are extracted from the image and contain different types of information. Three different techniques will be investigated and evaluated according to several evaluation methods. The methods that are used here are decision tree, random forest, & binary logistic regression. These methodologies have been explained in section 2 and all work with

numerical and categorical input. The input values of all these methods are features that have been extracted from the images and the output values of the convolutional neural network. This means that the output values of the neural network serves as an input for the feature integration methods. This leads to a general design lay-out which serves as a guideline throughout the entire modeling process.

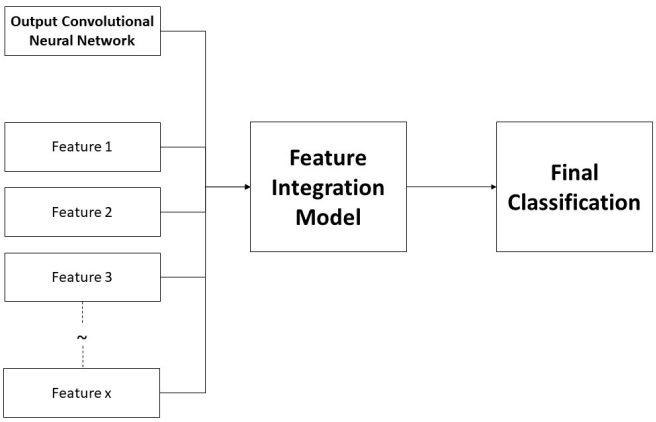

Figure 3.3: General lay-out experimental setup of feature integration

As can be seen in figure 3.3, the decision has been made to treat the output values of the convolutional neural network as input values for the different feature integration methods. This is a design choice and there are several arguments for why the lay-out has been designed this way. First of all, by treating the output values of the convolutional neural network as input values similar to the values of the other features, no bias in which features have more weight or influence on the final classification is created. The different techniques will be tested and evaluated and the results of the analysis will show which input values have more impact on the result. Secondly, the lay-out is consistent and therefore implementable for the different techniques. Moreover, other techniques that will not be tested can probably make use this structure as well. Finally, other lay-out structures are less desired for other reasons. This can be explained with the example of a simultaneous structure as opposed to the presented sequential structure. In this simultaneous structure the feature integration method is separated from the neural network and trained apart, after which the combination of the two methods leads to a final classification. A major disadvantage of this structure is the difficulty to make a weighted decision about which of the two method should get more emphasis and how the final classification is made. In the presented sequential structure this problem does not exist as only one method exists that performs the final classification, which is the feature integration method. Conclusively, these arguments have led to the lay-out of the experimental set-up as presented in figure 3.3

3.4.2 Evaluation of the different models

The different models share similar evaluation methods, but also differ in some of these methods. This is because some of the models require model-specific evaluation methods to measure its quality. However, the general classification performance of the models should be compared as well, which should be done by evaluation methods that can be used for all the different models. Therefore, a distinction has been made in model specific evaluation methods and evaluation methods that measure the classification performance. Both types of evaluation methods are required to make a good evaluation on the models.

In order to retrieve the best model amongst different generated models, some model-specific evaluation methods should be treated. First of all, when evaluating the performance of a decision tree, the Gini-impurity will be used to evaluate the quality of each decision. This is a good measure to use when establishing the depth of the tree. The Gini-impurity has been explained in more detail in section 2.5. Secondly, the performance of a regression analysis requires some specific evaluation methods as well. For example, the p-value of the overall model and the different parameters gives an indication of the significance of the model. If the p-values of an independent variable is lower than 0.05, the probability that variable on the classification is lower than 5%. If the overall p-value of the model is lower than 0.05, the probability that a new sample will be classified randomly is lower than 5%. Finally, for the comparison of the different regression models, the AIC-score will be used. This score in itself is meaningless, but it can be used to compare models that use the same data. Then, a lower AIC-score indicates a better model. This measure is recorded in the sections for the binary logistic regression.

In order to compare the different methods that integrate the additional features, the classification performance of the output classes will be compared in the same way as for the convolutional neural network. Firstly, a confusion matrix will be constructed of the actual and predicted output classes. This should give a graphical overview of the performance of the model. Afterwards, the overall accuracy will be measured in order to get a general idea of the performance of the convolutional neural network. However, this measurement is not accurate enough in itself to establish a correct opinion on the quality of the model. A detailed look into the recall and precision of the different categories should indicate whether the model could be overfitted or underfitted. The combination of these evaluation methods should give enough information to make an educated evaluation about the performance of the model. A detailed explanation on these evaluation methods has been provided in section 2.5.

4 Image Classification

This section is dedicated to report the results of the first part of the project that embraces the initial classification of non-melanoma skin cancer based on image data. First, the architecture and the results of the CNN that distinguishes two classes is presented, after which the architecture and the results of a binary CNN, which is able to detect NMSC in general, will be discussed. Therefore, the focus of this section is mostly on the evaluation part of the CRISP-DM methodology in the first part of the project.

4.1 Convolutional neural network with 6 classes

4.1.1 Network architecture

Multiple different models have been tested in order to discover which architecture gives the most accurate and consistent results. In order to do this, various parameters have been altered, such as the number of convolutional layers, the number of fully connected layers and the amount of nodes per layer. An elaborate overview of these parameters and the different options that have been considered has been described in section 3.3. Compared to the other tested models, the convolutional neural network in figure 4.1 is the best performing network and is designed as follows:

- Convolutional layer of 48 nodes
- Pooling layer
- Convolutional layer of 80 nodes
- Pooling layer
- Convolutional layer of 80 nodes
- Pooling layer
- Flattening layer
- Dropout layer (20% dropped)
- Dense layer of 120 nodes
- Dense output layer of 6 nodes with a SoftMax activation function

The architecture of the neural network has been illustrated in 4.1. The vectors with numbers in each layer show the dimensions of the image as it flows through the network, which change massively flowing throughout the process. After the input layer, the first convolutional layer of 48 nodes has been put. The first convolution takes place here, which reduces the image size from 64 by 64 by 3, to 62 by 62 by 48. The increase in the third from 3 to 48 is caused by the layer having a total of 48 nodes. The pooling layers, reduce the size of the image by a factor 4 to 31 by 31 by 3, as a stride of 2 by 2 moves over the neural network to select the input values that are the highest. This result is repeated twice with convolutional layers of 80 nodes leading to an input size of 6 by 6 by 80 for the flattening layer. This flattening layer stretches the image to one vector of length 2880, but does not have weights itself and therefore is only implemented to stretch the results, so it can enter the fully connected layer. This fully connected layer consists of 120 nodes and is

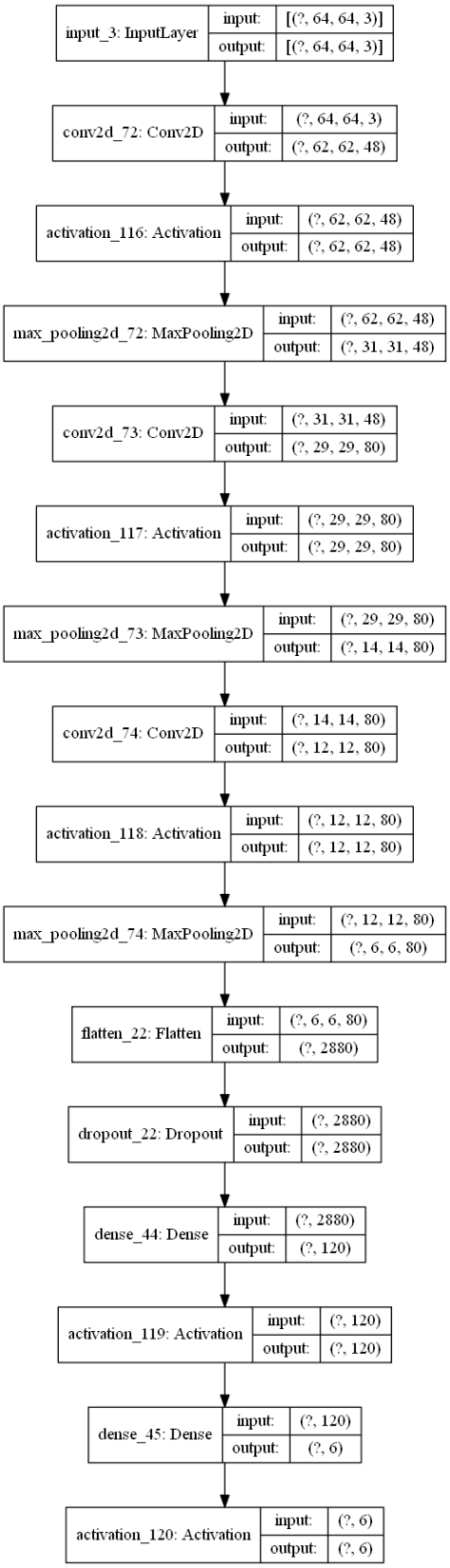

Figure 4.1: Architecture of the Convolutional Neural Network with 6 classes

activated by a ReLU function, which is also the case for the convolutional layers. The final layer is also a dense layer and translates the values to a probability score per category. This probability score determines the final classification. The category that has the highest prediction score will be selected as the ultimate classification. The evaluation of the performance of the CNN will be based on how well the model is able to do this ultimate classification. The prediction scores themselves will serve as input for the second part of the project. The results of this second part have been explained more elaborately in chapter 5.

4.1.2 Results

The first thing that is being evaluated is how well the CNN performs on the test set in general. As explained several evaluation methods have been selected upfront to determine the performance of the Convolutional Neural Network. Precision & recall are the most reliable results from a scientific point of view. Precision and recall can be combined into an F1-score together. More information on the evaluation methods can be found in section 2.5. These scores are retrieved from the confusion matrix, which provides an overview of the total amount of correct classifications.

Table 4.1: Confusion Matrix of the CNN

	Predicted labels					
True Labels	Verrucae	ID	Naevi	AK	BCC	SCC
Verrucae	19	7	4	5	6	3
ID	1	22	0	12	3	0
Naevi	2	0	30	6	1	0
AK	0	3	0	47	5	0
BCC	2	2	1	8	38	8
SCC	3	0	1	8	7	23

Firstly, what clearly comes forward from the confusion matrix is that the most often predicted class is always the correct class. This is an important result, because this shows that the CNN is able to make clear distinctions. Secondly, the most common mistake of the model appears to be an Inflammatory Dermatosis classified as an AK. This makes sense because the color and the scaliness of certain Inflammatory Dermatoses have similar features as AKs. Therefore, the expectation that similar patterns exist between Inflammatory Dermatoses and AKs. Finally, Naevi appear to have been predicted very well, both in terms of precision and recall. The recall of AK seems to be very high as well. A classification report of this confusion matrix can provide more detailed information.

Table 4.2: Classification report of the CNN

	Precision	Recall	F1 score
Verrucae	0.70	0.43	0.54
ID	0.65	0.58	0.61
Naevi	0.83	0.77	0.80
AK	0.55	0.85	0.67
BCC	0.63	0.64	0.64
SCC	0.68	0.55	0.61
Weighted Average	0.66	0.65	0.65

What comes forward in table 4.2 the model has a very high recall on Naevi and AK, which has been concluded from table 4.1. An explanation for the good performance on Naevi is the minor differences between the pictures of the Naevi. Most of the Naevi are lesions with a very consistent outline, which gives the model a lot of input to analyze the Naevi very well. An explanation for the high recall for AK is that the model appears to be a little overfitted on AK. This comes forward in the low value of precision, which is only 0.55. The opposite has happened for verrucae, which has a high precision, but a low recall. This can be an indication for the model being underfitted on verrucae. This is likely, since the database with images of verrucae is the smallest. These results come forward in the F1 score of the different categories. Here, AK does not have an F1 score much higher than the average F1 score of the model. Besides, the F1 score gives the indication that the Naevi are the best classified category of the model, scoring well on both precision and recall. Overall, the model has an overall accuracy of 64.6%

4.1.3 Merging classifications

In order to put a different perspective on the results, the confusion matrix can be merged into a confusion matrix with binary classification. This classification will be either skin cancer or no skin cancer. The distinction between skin cancer and no skin cancer is the most important one for MohsA, mainly because that impacts the decision on whether an appointment on the clinic is required.

This aggregation means that several incorrect classification become correct in this evaluation. For example, a verruca that has been wrongly classified as a Naevi, will become a true positive in this evaluation, because both verrucae and Naevi are no skin cancers. Similarly, errors between AK, BCC and SCC disappear and will all count as a correct skin cancer classification.

Table 4.3: Classification report of the CNN with merged classes

	Precision	Recall	F1 score
No skin cancer	0.876	0.702	0.780
Skin cancer	0.800	0.923	0.857

Table 4.3 presents an overview of the different performance measures. In section 4.1.2 it was discussed that the model is slightly overfitted on AK, and underfitted on Verrucae. This is confirmed by table 4.3. The difference in recall between skin cancer and no skin cancer is very significant in favor of skin cancer, whereas the precision of no skin cancer is higher. This is mainly because the value "skin cancer" has been predicted way more often than should be. Nevertheless, this is not a bad result for two important reasons. Firstly, for MohsA it is more important to make sure as few skin cancers are missed as possible. Therefore, it is worse to miss a skin cancer because it was labeled wrongly than to judge an innocent lesion as a skin cancer. Although, this would still mean longer waiting times at the clinic, which was the main goal of writing this thesis, the costs of a skin cancer that is spotted too late or not at all are way higher due to longer, more intensive treatments and higher treatment costs. Secondly, the CNN only serves as an input for the second part of the project. Therefore, a perfect model is not required, since this initial input will potentially be improved. The overall accuracy of this model is 82.67%.

4.2 Convolutional Neural Network with binary classification

Other purposes of the neural network have been investigated as well. The distinction between the 6 categories can be generalized into a distinction between 2 categories: skin cancer and no skin cancer. The goal of executing this binary classification is to investigate to which extent the network is able to make this distinction and whether it is better or worse than the aggregated version of the original CNN where the classes have been trained separately, as discussed in section 4.1.3.

4.2.1 Differences & similarities with original CNN

For the biggest part the design of the CNN with binary classification is similar to the CNN with categorical classification. Thus, the binary network consists of three convolutional layers, one 20% dropout layer, one dense layer, and 80 nodes per layer, except from the first layer which also consists of 48 nodes per layer. Moreover, the network uses an Adam optimizer.
However, the binary classification problem requires some adaptations of the original CNN as well. Firstly, the number of output classes changes from 6 to 2. This means that the output layer only has one node producing either 0 or 1 as output value, with 0 not being a skin cancer and a 1 being one. Therefore, the data is being labeled in a slightly different way. Moreover, this merger means that the results of this network can not compared one on one with the results of the multi-class CNN. This is because the probability of guessing the correct class is 50% in the binary model, while it is 16.67% in the multi-class neural network. This means that comparing the models one on one by accuracy, precision, recall & F1-score does not give a fair indication of the performance of the model. Another difference between the models is the use of a Sigmoid activation function in the output layer instead of a SoftMax function. This is because a Sigmoid function is a derivative of the SoftMax classifier and is simpler, which means the computation time is lower. Therefore, Sigmoid functions are almost always used in the final dense layer of a binary CNN instead of SoftMax functions. Finally, the binary cross entropy loss function is used instead of sparse categorical cross-entropy, since a binary classification problem requires a binary

loss function. The visual architecture of the binary CNN looks exactly the same as in figure 4.1, except for the final output layer.

4.2.2 Results

Likewise to the results of the categorical CNN, the performance of the binary CNN can be evaluated through a confusion matrix and corresponding classification report.

Table 4.4: Confusion Matrix of the binary CNN

	Predicted No Skin Cancer	Predicted Skin Cancer
Actual No Skin Cancer	92	29
Actual Skin Cancer	17	139

Table 4.5: Classification report of the binary CNN

	Precision	Recall	F1 score
No skin cancer	0.844	0.760	0.800
Skin cancer	0.827	0.891	0.858

The confusion report and the classification report are presented in table 4.4 & 4.5. The tables show an especially good result on the recall of skin cancer, which is 89%. Although the recall on no skin cancer is a little lower at 76%, the differences between the recall values are lower than for the CNN with merged classes, which were 92.3% and 70.2% respectively. Therefore, the binary CNN seems to be able to make a more balanced prediction than the CNN with 6 classes. In order to make a good evaluation on which of the two CNNs performs better, the results need to be compared on the different evaluation methods. The results of the two models are compared on recall, precision and F1-score of skin cancer, since that is the most important class to detect, and overall model accuracy.

Table 4.6: Comparing two CNN approaches

	Multi-class CNN	Binary CNN
Precision Skin Cancer	0.8	**0.827**
Recall Skin Cancer	**0.923**	0.891
F1 Skin Cancer	0.857	**0.858**
Overall Accuracy	0.826	**0.834**

As can be seen in table 4.6, the binary CNN outperforms the multi-class neural network on most aspects, except for the recall of skin cancer, on which the multi-class CNN performs slightly better. However, since the binary CNN scores higher on the other evaluation measures, the conclusion arises that the binary CNN is more balanced, less overfitted on skin cancer and is

better suitable for making a correct classification among skin cancer and no skin cancer in terms of overall accuracy. Nevertheless, the results are not extremely different and both approaches could be suitable. However, the binary CNN seems to produce slightly more robust results. For that reason, the results of the binary CNN will be used as input for the binary random forest.

5 Integration of Additional Features

This section is dedicated to the improvement of the results of the convolutional neural networks as described in chapter 4. The results of these neural networks will serve as an input value in the different additional techniques that are described in this chapter. Added to these input values are features that have been manually retrieved from the images themselves. The described methods are the decision tree, random forests, & binary logistic regression. The decision tree and the random forest will be used to make a prediction between the six different classes using the input values of the multi-class CNN. The binary logistic regression and again the random forest will be used to make a binary classification between skin cancer and no skin cancer. Moreover, the results of the multi-class prediction models will be merged into a binary classification model and will be compared with the models that used a binary input. The results of these models are reported in this chapter.

5.1 Decision Tree

Before entering the random forest, several insights can be retrieved from looking at a single decision tree. First of all, it provides more insights in the way a random forest is structured, as a random forest is an aggregation of multiple decision trees. Contrary to the random forest, a decision tree is easy to visualize as its structure can be displayed in one single image, which is not possible for a random forest. Secondly, although a single tree does not have a high robustness, some information can be retrieved with regard to the accuracy of a decision tree. Comparing the accuracy of one decision tree with an entire random forest provides insights in to which extent a random forest actually adds value compared to a single decision tree.
Not a lot of parameters need to be altered before running the decision tree model. Only the maximum depth of the decision tree, which relates to the amount of split decisions being made before the final decision, has been set to eight, which is similar to the depth of the random forest. An image of the generated decision tree can be found in appendix A.

Table 5.1: Table of the relative importance of each feature in the decision tree

Feature	Relative importance
Prediction: Naevi	0.21
Prediction: BCC	0.18
Prediction: SCC	0.18
Prediction: AK	0.16
Prediction: Verrucae	0.13
Prediction: ID	0.03
Location	0.03
Shininess	0.03
Color	0.01
Red Edge	0.01
Elevation	0.01
Scaliness	0.00
White Inside	0.00
Blood Clot	0.00

The most important conclusion that can be drawn from table 5.1, is that when it comes to one single decision tree, not all the factors can be taken into account. This leads to Scaliness, White Inside and Blood Clot contributing 0% to the final classification. Moreover, the factors that actually contribute mostly are the prediction values from the CNN for the 6 different categories, although surprisingly, the prediction score for ID does not contribute as much as the other factors. This can be explained by the randomness of a single decision tree. The expectation is that a random forest, which is designed to level out this randomness, would have more balance between the different prediction scores.

Table 5.2: Confusion Matrix of the Decision Tree

	Predicted labels					
True Labels	Verrucae	ID	Naevi	AK	BCC	SCC
Verrucae	22	5	5	4	7	1
ID	3	19	0	13	3	0
Naevi	6	0	30	3	0	0
AK	2	2	0	46	5	0
BCC	4	3	2	5	33	11
SCC	6	1	0	8	4	23

Table 5.2 shows high discrepancies between the categories. Some categories like Verrucae and ID have not been predicted very accurately, whereas categories such as AK & Naevi are very often predicted correctly. This relates to the results of the CNN, where AK & Naevi were also the best predicted categories, while Verrucae had a low score on both precision and recall.

Table 5.3: Classification report of the Decision Tree

	Precision	Recall	F1 score
Verrucae	0.51	0.50	0.51
Inflammatory Dermatoses	0.63	0.50	0.56
Naevi	0.81	0.77	0.79
AK	0.57	0.84	0.68
BCC	0.63	0.56	0.59
SCC	0.66	0.55	0.60
Weighted Average	0.63	0.62	0.62

Table 5.3 shows the results from the different evaluation methods of the decision tree. The table confirms the preliminary conclusions drawn from the confusion matrix. Firstly, the recall of the Verrucae and ID are very low. This result is very similar to the results of the CNN, which has a big impact on this decision tree. Secondly, Naevi and AK have been predicted relatively accurately. This result aligns with the results from the CNN as well, where these categories performed very similar.

5.2 Random Forest - Multi-class classification

In this section the results of the random forest will be discussed. The results will be structured the same way as in section 5.1. Several parameters for the random forest are being set, before the model is being developed. These parameters have been tweaked and tested and the best possible results have been taken. First of all, the random forest will consist of 100 decision trees. This number is not too high to cause the random forest to be overfitted. However, 100 decision trees is high enough to generate a robust answer without a too high random factor. Secondly, the maximal depth of a tree will be 8, equal to the decision tree in section 5.1. Similar to the reasoning for the number of tree, this number should not be too high in order to avoid over-training, but not too low to remain robust answers.

Table 5.4: Confusion Matrix of the Random Forest

	Predicted labels					
True Labels	Verrucae	ID	Naevi	AK	BCC	SCC
Verrucae	25	6	5	2	4	2
ID	2	21	0	11	4	0
Naevi	5	0	31	3	0	0
AK	0	4	0	44	6	1
BCC	4	2	1	5	39	8
SCC	4	0	1	8	2	27

Firstly, what comes forward from the confusion matrix, and what also came forward from the CNN described in section 4.1.2, is that the most often predicted class is always the correct class. This

is a positive result, because it confirms that the random forest is able to make a clear distinction between the different categories based on the input features. Secondly, where in the CNN, the most common confusion was an inflammatory dermatosis being classified as an AK, the confusion of an SCC being classified as an AK is now equally often diagnosed with a total of 8 times. This is mainly due to the reduction of the inflammatory dermatosis being classified as an AK. This shows that the RF appears to correct for common confusions of the CNN, which would be a positive result.

Table 5.5: Classification report of the Random Forest

	Precision	Recall	F1 score
Verrucae	0.625	0.568	0.595
Inflammatory Dermatoses	0.636	0.553	0.592
Naevi	0.816	0.795	0.805
AK	0.603	0.800	0.688
BCC	0.709	0.661	0.684
SCC	0.711	0.643	0.675
Weighted Average	0.680	0.675	0.674

Table 4.2 provides an overview of the evaluation measures for the random forests. Compared to the CNN, the overall accuracy has increased from 64.6% to 67.5%, which is an improvement. The F1 score of all classes, except ID, has increased in the random forest. This is mainly caused by the overall increase in recall of the most classes. Only the recall of AK has dropped. However, as AK seemed to be a little overfitted on the CNN, the lower score on recall is not a problem as F1 has increased. Overall, the difference in performance on the different classes appears to be better and the difference between the best performing class, Naevi, and the worst performing class, Verrucae, has decreased on both recall and F1 score. Conclusively, the random forest actually appears to improve the initial result from the CNN and reduces its variance. Therefore, it is interesting to analyze the factors that contribute most to the result of the random forest.

Table 5.6: Table of the relative importance of each feature in the random forest

Feature	Relative importance
Prediction: Inflammatory Dermatoses	0.15
Prediction: Verrucae	0.15
Prediction: SCC	0.13
Prediction: Naevi	0.13
Prediction: AK	0.11
Prediction: BCC	0.11
Color	0.08
White Inside	0.03
Location	0.02
Shininess	0.02
Red Edge	0.02
Elevation	0.02
Blood Clot	0.02
Scaliness	0.01
Skin Type	0.00

As can be seen in table 5.6, the factors that contribute the most to the random forest are the values that already come from the CNN. Without exception, these six predictive variables are the most contributing variables by making up 78% of the total prediction. This confirms the hypothesis from section 5.1 that these factors would have the highest contribution for the classification in the random forest. The other 22% is mostly within the color of the lesion at 89%, while the other factors share a more equal contribution. Interestingly, skin type does not contribute at all to the results of the random forest. This makes sense, because as explained in section 3.2.3.2, the data set is too imbalanced on this subject to make correct predictions with regard to this feature.

5.2.1 Random Forest without prediction values

For comparison, a random forest has been produced that does not make use of the prediction scores. Since the other features have been included in the random forest, the random forest is almost similar to the originally used one except from the absence of the prediction scores. Two main reasons exist to build this random forest. Firstly, if the results from this random forest are similar or better than the random forest with the prediction scores of the CNN, we can conclude that the prediction scores have limited impact on the correct classification of the different images compared to the extracted features. Secondly, if this random forest shows very poor results, we know that the impact the features can have on the results through a random forest is quite minimal.

Table 5.7: Classification report of the Random Forest without prediction scores

	Precision	Recall	F1 score
Verrucae	0.338	0.568	0.430
Inflammatory Dermatoses	0.581	0.553	0.522
Naevi	0.727	0.615	0.667
AK	0.615	0.436	0.511
BCC	0.436	0.441	0.495
SCC	0.711	0.643	0.581
Weighted Average	0.559	0.523	0.529

The classification report in table 5.7 shows the results of this random forest in terms of recall, precision and F1-score. Almost all lesions have not been predicted very well. Especially verrucae, which only has a precision of 33.8% and an F1-score of 43%. Only Naevi are performing fine compared to the other categories, but both the scores on precision and recall are way lower when the prediction scores are included. Nevertheless, the classification is not completely random and the average recall is higher than 50%. The average accuracy is 52.3% over the 6 categories. This shows that the prediction values themselves have added value to the random forest and that it improves the result that would have been generated from the features retrieved from the images. Besides, it shows that the features themselves contain enough information to make some sort of prediction. However, the result is clearly not good enough to make reliable predictions.

5.2.2 Merging classifications

Similarly to section 4.1.3, an interesting view can be provided on when the classes are merged and a binary classification will be provided between skin cancer and no skin cancer. As explained previously, the distinction between skin cancer and no skin cancer is the most important one for MohsA, mainly because that impacts the decision on whether an appointment on the clinic is required.
This aggregation means that several incorrect classifications become correct in this evaluation. For example, a verruca that has been wrongly classified as a naevi, will become a true positive in this evaluation, because both verrucae and Naevi are no skin cancers. Similarly, errors between AK, BCC and SCC disappear and will all count as a correct skin cancer classification.

Table 5.8: Classification report of the Random Forest with merged classes

	Precision	Recall	F1 score
No skin cancer	0.855	0.783	0.819
Skin cancer	0.843	0.897	0.870

The results in table 5.8 can be compared to the results in section 4.1.3. First of all, the recall of no skin cancer has clearly improved, as it was 0.70 after merging the classes of the CNN and is now

0.785. This is a positive result that indicates an improvement in performance on no skin cancer. Secondly, the recall for skin cancer has stayed high at 89.7%. Overall, the F1 scores have improved for both categories, which is an indication that the model has improved the initial results from the CNN.

5.3 Random Forest - Binary classification

Keeping in mind the current results of the random forest for 6 classes and the aggregation that has been made in section 5.2.2, it is interesting to compare this result to a random forest with binary classification. Making use of the predictions of the binary CNN from section 4.2, a random forest can be constructed that makes a distinction between skin cancer and no skin cancer.
The design of the binary random forest is similar to the random forest with 6 output values. The total number of trees in the random forest is 100, and each tree has a maximum depth of 8. However, some differences are present in terms of input values. Instead of the input of the multi-class CNN, the binary CNN has been used as a source of input. This implies that instead of a single prediction score for one of 6 output classes, just one prediction score will serve as input from the CNN. This prediction score is a value between 0 and 1, where 0 indicates no skin cancer and 1 indicates skin cancer.

Table 5.9: Classification report of the binary Random Forest

	Precision	Recall	F1 score
No skin cancer	0.847	0.868	0.857
Skin cancer	0.895	0.878	0.887

The classification report in table 5.9 shows some fine results. First of all, the variance between the different output variables in terms of precision and recall appear to be low, indicating that the model is not very overfitted on one particular class. Although the results on skin cancer are better than on no skin cancer with an F1 score of 0.887, the F1 score for no skin cancer is also 0.857. The overall model accuracy is 87.4% and is an improvement from the binary CNN which had an accuracy of 83.4%. Conclusively, the binary random forest appears to be a nice measure for the classification of skin cancer.

Table 5.10: Table of the relative importance of each feature in the random forest

Feature	Relative importance
Prediction	0.54
Location	0.13
Color	0.12
Blood Clot	0.09
Scaliness	0.04
White Inside	0.03
Shininess	0.02
Elevation	0.01
Red Edge	0.01
Skin Type	0.01

Table 5.10 shows the relative importance of each feature for the classification of the random forest and some interesting conclusions can be drawn, especially compared to the random forest with 6 classes. First of all, the prediction score is far less important in this section than it was in the random forest with 6 classes. In the binary classification it makes up for 54% of the prediction, whereas this was 78% in the multi-class random forest. Secondly, the gap this leaves to the contribution of other variables is mostly filled by the variables 'Blood clot' and 'Location'. 'Blood Clot' has a relative importance of 9% here compared to 2% for the multi-class random forest. 'Location' even has an importance of 13% compared to the 2% in the multi-class random forest. Therefore, the conclusion arises that the location of the lesion and the presence of a blood clot are important features to assess whether a lesion is skin cancer or not, but that those features are less useful when an exact classification is required.

Another interesting viewing point is to look what would make a better distinction between skin cancer and no skin cancer: a random forest trained on 6 different classes, after which the results are merged into two categories, or a random forest that is trained on only 2 classes: skin cancer and no skin cancer. The results of the two models are compared on recall, precision and F1-score of skin cancer, since that is the most important class to detect, and overall model accuracy.

Table 5.11: Comparing two Random Forest approaches

	Multi-class Random Forest	Binary Random Forest
Precision Skin Cancer	0.843	**0.895**
Recall Skin Cancer	**0.897**	0.878
F1 Skin Cancer	0.870	**0.887**
Overall Accuracy	0.848	**0.874**

As can be seen in table 5.11, the binary random forest outperforms the multi-class random forest on most aspects. Only on the recall of skin cancer, the multi-class random forest performs slightly better. However, since the binary random forest scores higher on the other evaluation measures, the conclusion arises that the binary random forest is more balanced, not clearly overfitted and is better suitable for making a correct classification among skin cancer and no skin cancer.

5.4 Binary logistic regression

The final method that is considered for the integration of additional features is the binary logistic regression. In a binary logistic regression model only a distinction between two classes can be made. Therefore, this methods will be applied to make a distinction between images that are non-melanoma skin cancer and the other lesions. Binary logistic regression will also be used to make a distinction between an individual class and the other lesions. The classes on which this technique is applied are actinic keratosis (AK), basal cell carcinoma (BCC) and squamous cell carcinoma (SCC).

5.4.1 Binary logistic regression - Non-melanoma skin cancer

First of all, binary logistic regression is used to detect non-melanoma skin cancer amongst the other lesions. Therefore, AK, BCC & SCC are labeled as skin cancer, whereas the other lesions are labeled as no skin cancer. The output of the formula is a number between 0 and 1, with 1 being 100% certainty about the image being skin cancer, while 0 means no skin cancer. Therefore, if the output number is rounded to the nearest integer, a prediction can be made. The total of predictions results in the following confusion matrix:

Table 5.12: Confusion Matrix of the binary logistic regression model

	Predicted Skin Cancer	Predicted No Skin Cancer
Actual Skin Cancer	145	11
Actual No Skin Cancer	24	97

The following classification report corresponds with the confusion matrix in table 5.12.

Table 5.13: Classification report of the binary logistic regression model

	Precision	Recall	F1 score
Skin cancer	0.858	0.929	0.892
No skin cancer	0.898	0.802	0.847

Several conclusions can be drawn from the classification report in table 5.13. First of all, the precision and recall appear to be more balanced than the original CNN, but still slightly overfitted towards skin cancer. This is also reflected in the F1 score, which is very high for skin cancer at 89.2%. Besides, for the detection of skin cancer, the scores are the highest of all the tested models.

This comes forward in the model accuracy as well, which is 87.4%

The results from this binary linear regression model are based on a rounded prediction number. A probability > 0.5 leads to the prediction skin cancer and a probability <0.5 indicates no skin cancer. Therefore, the expectation arises that the most errors occur around 0.5. Besides, it is interesting to see from what point the model almost makes no errors anymore

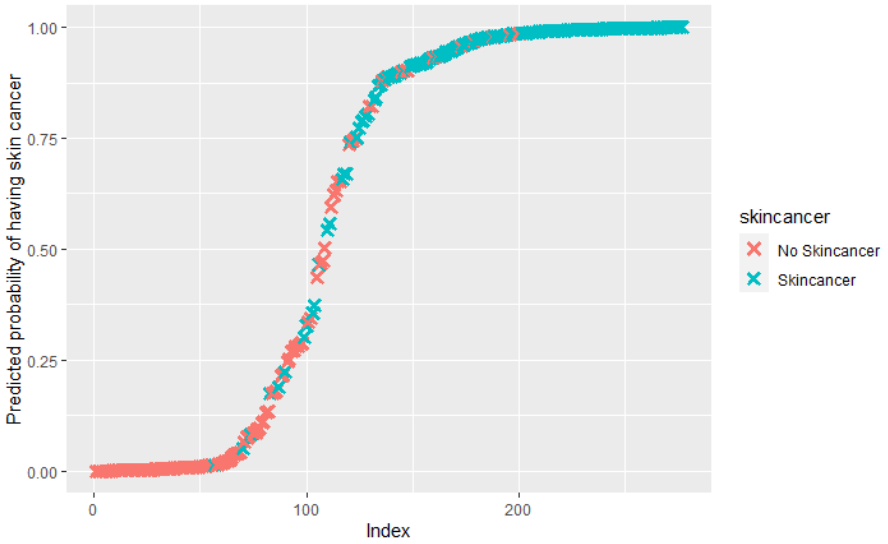

Figure 5.1: Predicted probabilities of having skin cancer by the binary logistic regression model

The images are presented in figure 5.1 on the x-axis and have been sorted on the predicted probability of having skin cancer, with the lowest index having the least probability of having skin cancer. The actual predicted probability for skin cancer is predicted on the y-axis. The graph shows that the most errors occur in the range between 0.15 and 0.75, which is logical since the model is less sure on whether it is skin cancer or not. However, the model appears to have a very small error rate when the predicted probability is higher than 0.75. Zooming in on the lesions with a probability higher than 0.75 leads to the following graph.

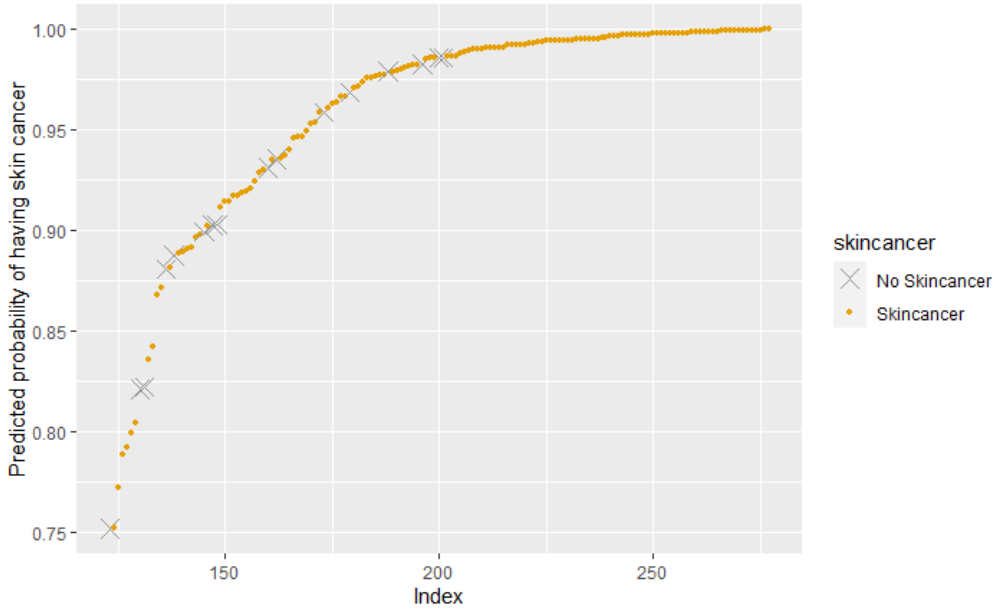

Figure 5.2: Images with a predicted probability > 0.75 by the binary logistic regression model

Figure 5.2 shows that not a lot of errors seem to be made at a probability score above 0.75. The graph presents a total of 138 samples from the test set of which the probability score from the binary logistic regression model is higher than 0.75. From these samples, only 8 are actually no skin cancer. This means if the model predicts the probability for skin cancer to be higher than 75%, 94.2% of the time this is actually correct. This is a result that can impact the practical implications of the model. For example, if a doctor uses this model to verify his own suspicions about a lesion and the model gives a score that is higher than 75%, the doctor will know that the lesions is a type of skin cancer with 94.2% certainty. This is a useful result, that can support doctors in their decision making.

Table 5.14: Summary of the logistic regression model for the prediction of skin cancer

Independent variables	Location: Hand, Head, Back Color: Yellow, Red, Light Red, White Shininess White Inside Blood Clot Elevation
Residual Deviance	617.12 on 1944 degrees of freedom
AIC	701.12
p-value < 0.05?	0 (significant)

Table 5.14 provides a detailed summary of the regression model. First of all, the significant independent variables have been described. Interestingly, the prediction values for the different

types of skin cancer are not significant. The p-values of these values is 0.187, which does not indicate significance. The other mentioned variables do have significant impact on the output of the logistic regression model. The residual deviance is 617.12 on 1944 degrees of freedom, which indicates a p-value for the model of 0. This means that the model is significant.

5.4.2 Binary logistic regression - AK

Beside the binary distinction between skin cancer and no skin cancer, binary logistic regression can be applied in order to detect the different types of skin cancer separately as well. First, AK will be the skin cancer to be detected among the other lesions. Therefore, the data has been divided in the images that are AK and the ones that are not AK.

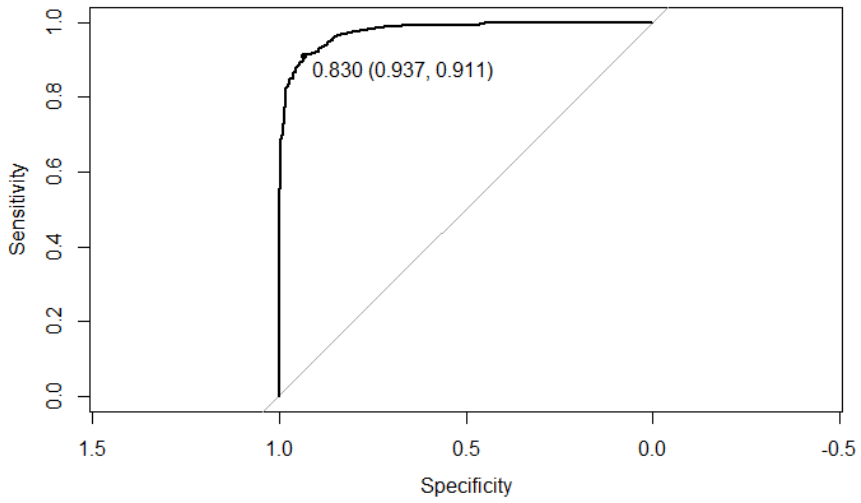

Figure 5.3: ROC-curve for the detection of AK with binary logistic regression

Figure 5.3 shows the ROC curve for the detection of AK. The threshold for distinction between AK and other lesions for optimal sensibility and specificity is at a probability value of 0.830, which in this particular regression means that if the probability value from the logistic function is lower than 0.830, the lesion will be classified as a AK. This leads to a sensibility of 91.1% at a specificity of 93.7% on the training set. On this threshold a confusion matrix can be made for the data in the test set.

Table 5.15: Confusion Matrix for binary logistic regression for detection of AK

	Predicted AK	Predicted Other Lesion
Actual AK	51	4
Actual Other Lesion	47	175

The confusion matrix in table 5.15 gives a sensitivity of 92.7% at a specificity of 78.8% at the threshold of 0.83. This appears to be a very good and suitable result that shows that AK can be

predicted with quite some confidence. Especially the high sensitivity shows that not a lot of AKs would be missed if the data is regressed this way.

Table 5.16: Summary of the logistic regression model for the prediction of AK

Independent variables	Location: Leg, Face, Head, Back, Nose Shininess White Inside Scaliness Elevation
Residual Deviance	572.47 on 1944 degrees of freedom
AIC	656.47
p-value < 0.05?	0 (significant)

Table 5.16 provides a detailed summary of the regression model. First of all, the significant independent variables have been described. Interestingly, the prediction values for the different types of skin cancer are not significant. The p-values of these values is 0.116, which is not enough to indicate significance. The other mentioned variables do have significant impact on the output of the logistic regression model. The residual deviance is 527.47 on 1944 degrees of freedom, which indicates a p-value for the model of 0. This means that the model is significant.

5.4.3 Binary logistic regression - BCC

The same thing that has been done in section 5.4.2 can be executed for the detection of BCC. In order to receive the threshold for the optimal balance between sensitivity and specificity, the ROC-curve will be plotted for the detection of the threshold for the probability of BCC.

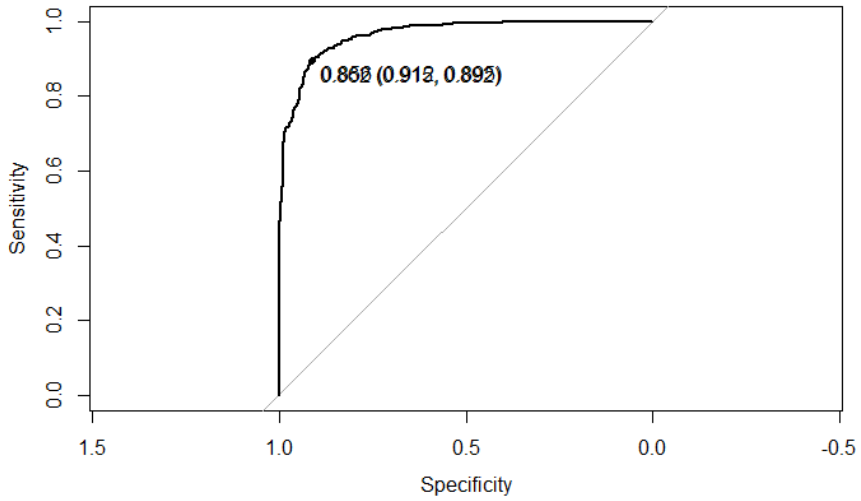

Figure 5.4: ROC-curve for the detection of BCC with binary logistic regression

Figure 5.4 shows the ROC curve for the detection of BCC. The threshold for distinction between BCC and other lesions for optimal sensibility and specificity is at a probability value of 0.856, which in this particular regression means that if the output value from the logistic function is lower than 0.856, the lesion will be classified as a BCC. This leads to a sensibility of 89.2% at a specificity of 91.5% on the training set. On this threshold a confusion matrix can be made for the data in the test set.

Table 5.17: Confusion Matrix for binary logistic regression for detection of BCC

	Predicted BCC	Predicted Other Lesion
Actual BCC	49	10
Actual Other Lesion	29	189

The confusion matrix in table 5.17 gives a sensitivity of 83.1% at a specificity of 86.7% at the threshold of 0.856. Similar to the results for the logistic regression of AK, this appears to be a suitable results that shows that BCC can be predicted with quite some confidence, despite the sensitivity being lower than for AK. However, this regression shows a specificity that is higher than at AK. Moreover, the specificity and sensitivity are more aligned implicating a robust result.

Table 5.18: Summary of the logistic regression model for the prediction of BCC

Independent variables	Location: Face Shininess White Inside Red Edge Blood Clot
Residual Deviance	427.52 on 1944 degrees of freedom
AIC	511.52
p-value < 0.05?	0 (significant)

Table 5.18 provides a detailed summary of the regression model. First of all, the significant independent variables have been described. Interestingly, the prediction values for the different types of skin cancer are not significant. The p-values of these values is 0.715, which does not indicate significance. The other mentioned variables do have significant impact on the output of the logistic regression model. The residual deviance is 688.35 on 1944 degrees of freedom, which indicates a p-value for the model of 0. This means that the model is significant.

5.4.4 Binary logistic regression - SCC

Finally, the logistic regression will be used to detect SCC amongst the other lesions. First, the threshold needs to be retrieved for the most optimal result between sensitivity and specificity.

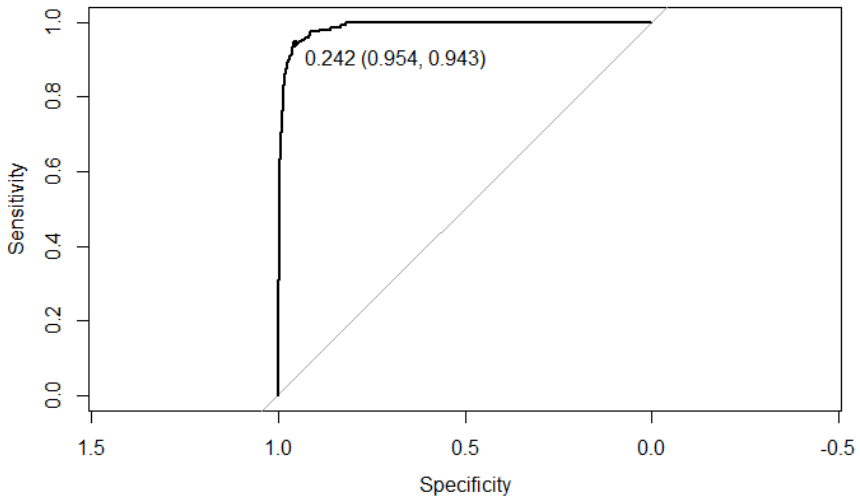

Figure 5.5: ROC-curve for the detection of SCC with binary logistic regression

Figure 5.5 shows the ROC curve for the detection of SCC. The threshold for distinction between SCC and other lesions for optimal sensibility and specificity is at a probability value of 0.242, which in this particular regression means that if the output value from the logistic function is higher than 0.242, the lesion will be classified as a SCC. This leads to a sensibility of 94.3% at a specificity of 95.4% on the training set, which is the highest sensitivity and specificity on the training set of the three lesions. However, to verify the quality of the model, only the results on the test set should be taken into account. On this threshold a confusion matrix can be made for the data in the test set.

Table 5.19: Confusion Matrix for binary logistic regression for detection of SCC

	Predicted SCC	Predicted Other Lesion
Actual SCC	31	11
Actual Other Lesion	11	224

The confusion matrix in table 5.19 gives a sensitivity of 73.8% at a specificity of 95.3% at the threshold of 0.242. Although this result is not extremely bad, a huge discrepancy exists between sensitivity and specificity. Clearly, the training set and the test set do not have the same optimum for the detection of SCC. Probably, a lower threshold for the detection of skin cancer would be more suitable, despite the fact that the optimum on the training set was on the 0.242.

Table 5.20: Summary of the logistic regression model for the prediction of SCC

Independent variables	Location: Hand, Head, Back Color: Yellow, Red, Light Red, White Shininess White Inside Blood Clot Elevation
Residual Deviance	617.12 on 1944 degrees of freedom
AIC	701.12
p-value < 0.05?	0 (significant)

Table 5.20 provides a detailed summary of the regression model. First of all, the significant independent variables have been described. Interestingly, the prediction values for the different types of skin cancer are not significant. The p-values of these values is 0.869, which does not indicate any significance. The other mentioned variables do have significant impact on the output of the logistic regression model. The residual deviance is 427.52 on 1944 degrees of freedom, which indicates a p-value for the model of 0. This means that the model is significant.

6 Discussion

This section is dedicated to the reflection upon the results of the different models. The results of the different models are compared and the model quality is assessed. First, results of the different multi-class classification methods have been compared.

Table 6.1: Comparison of different multi-class classifiers

Sensitivity	Verrucae	ID	Naevi	AK	BCC	SCC	Accuracy
CNN	43.2%	**57.9%**	76.9%	**85.4%**	64.4%	54.7%	64.6%
Decision Tree	50%	50%	77%	84%	56%	55%	62%
Random Forest	**56.8%**	55.3%	**79.5%**	80%	**66.1%**	**64.3%**	**67.5%**

Table 6.1 displays the different sensitivity scores for the four different classifying techniques that have been used to classify the six different categories: Verrucae, Inflammatory Dermatoses, Naevi, AK, BCC & SCC. It appears that the CNN and the random forest show the best performance on the different categories. Moreover, the random forest performs better than the CNN on most of the categories and on average accuracy. This is a logical result, because the random forest is implemented in order to improve the result of the CNN. Nevertheless, the CNN already provides a fine initial result, although the neural network seems to be biased towards AK and slightly underfitted on Verrucae. More data, while maintaining a balanced data-set could probably improve this issue.
The random forest without prediction scores is not a suitable approach. The sensitivity and accuracy scores mainly shows that the input from the CNN is an essential addition to the random forest. Besides, the single decision tree is also not suitable in terms of sensitivity, accuracy & stability. Out of the different options, just the random forest with prediction scores appears to improve the results from the CNN. Moreover, it seems to correct for some of the overfitted categories. However, the overall accuracy does not show enough improvement to prove significant added value. Addition of more categories, especially numerical ones, could show evidence for the hypothesis that the random forest is able to improve a CNN with 6 categories. Nevertheless, out of the tested methods for the classification of the 6 different categories, the random forest is the best method to enhance the image classification from the CNN.

Table 6.2: Comparison of different binary classifiers for the detection of non-melanoma skin cancer

	Sensitivity	Specificity	Accuracy
Binary CNN	89.1%	76.0%	83.4%
Multi-class CNN	92.3%	70.2%	82.6%
Binary RF	87.8%	**86.8%**	**87.4%**
Multi-class RF	89.7%	78.8%	84.8%
Logistic Regression	**92.9%**	80.1%	**87.4%**

Table 6.2 displays the different sensitivity and specificity scores for the different binary classific-

ation methods. Sensitivity and specificity are good evaluation methods in this section, because these methods are designed for indicating whether a certain disease, which is skin cancer in this case, is present or not. Several conclusions can be drawn from the table. First of all, the convolutional neural networks have a very high sensitivity compared to their specificity. Especially the imbalance for the multi-class CNN is clearly visible in this table. Secondly, the random forests are much more balanced, which can be seen in the higher specificity. This is especially the case for the binary random forest that shows the highest specificity. Moreover, the logistic regression shows very good results as well, especially in terms of sensitivity. Conclusively, in this research the binary random forest and the logistic regression are the best methods to detect skin cancer. The logistic regression is the best method to achieve a high sensitivity, whereas the binary random forest has the best balance. Both models score equally in terms of accuracy.

Table 6.3: Comparison of logistic regression for AK, BCC & SCC

	Sensitivity	Specificity
AK	92.7%	78.7%
BCC	83.1%	86.7%
SCC	73.8%	95.3%

Finally, the results of the individual regression of AK, BCC & SCC are displayed in table 6.3. Each of these categories have been individually regressed against all other categories. This leads towards some varying results. AK has the highest sensitivity at 92.7%. However, its specificity is the lowest at 78.8%. This can be an indication that the correct trade-off point has not been found. The results of SCC show the opposite imbalance. Despite a high specificity at 95.3%, the sensitivity is relatively low at 73.8%. In this case, the conclusion that the incorrect trade-off point might have been taken arises here. A more balanced result is displayed at BCC, which shows a 83.3% sensitivity at a 86.7% specificity. This is an indication that the correct trade-off point has been chosen. Moreover, other classifiers have not been able to detect BCC in the test set with such a high sensitivity. In conclusion, individual regression seems to be a good addition for the detection of the 3 different categories of skin cancer.

7 Conclusion

This final section is dedicated to provide a conclusion to this thesis. First of all, the main conclusions are summarized in detail. The elaborate summary aims to reflect upon the different research sub-question and the main research question. Secondly, an overview of the limitations in this project is provided. Thirdly, the practical implications and managerial insights are being provided, as multiple options exist to utilize these results. Finally, several suggestions for future research are summarized and described.

7.1 Summary

First of all, a convolutional neural network has been developed in order to distinguish between images of 6 different classes based on the pixel values in the images. The expectation was that convolutional neural networks are very suitable to make an initial classification of images, which is a conclusion based on the literature research. Considering the limited amount of data, the convolutional neural has made a good basic classification between the different categories. An accuracy of 64.6% between 6 classes is a results that can be improved, but within the scope of this project that would have been very difficult. Especially, the accuracy on Actinic Keratosis and Naevi is really good. Therefore, the conclusion is that the convolutional neural network is able to make a suitable distinction between the different classes. Moreover, both the binary neural network and the multi-class neural network proved to make an accurate distinction between skin cancer and no skin cancer, although this result showed that the neural networks were slightly over-fitted on the different skin cancers.

Secondly, out of all the tested models that combine the image classified data, several models performed better than other models, depending on the final goal. In order to make a distinction between the 6 categories Verrucae, ID, Naevi, SCC, BCC & AK, a decision tree and a random forest have been designed. The decision tree clearly was not capable enough in itself and did not improve the original results from the CNN. However, the random forest improved the initial classification from 64.4% to 67.5%, which is a very positive result. Not only does the average accuracy improve, the recall levels of the different categories is more balanced as well. Therefore, the random forest is the best option to make a distinction between 6 categories.

For the distinction between skin cancer and skin cancer, more suitable options have been compared. Binary logistic regression is one of the best performing models to detect skin cancer with the highest recall among all tested methods and a very high accuracy. However, in terms of specificity, the binary random forest outperforms the logistic regression. In terms of accuracy both models perform equally high at 87.4%. In total, both models prove to be very suitable for making correct distinctions between skin cancer and no skin cancer. Moreover, the binary random forest and the binary logistic regression both show a significantly more balanced result in terms of overfitting on skin cancer compared to both the binary CNN and the multi-class CNN, especially the binary random forest. Conclusively, these two models perform the best and show really promising

results in this area.

In order to make distinctions between one individual non-melanoma skin cancer and the other categories, some promising results have been achieved as well. First of all, AK can be detected by a logistic regression at a sensitivity of 92%, which is really high. Although the specificity is slightly lower, this is a promising result. Also for the classification of BCC a good result has been achieved with a sensitivity of 83.1% at a specificity of 86.7%. For SCC the sensitivity is much lower at 73.8%, but the specificity is 95.3%, which is highest specificity of the three categories, but also a clear indication that it is underfitted on SCC. An adjustment of the threshold within the regression formula could improve this, but this requires more data to successfully do this. Nevertheless, this result is promising as well.

However, in order to solve the initial problem, the developed model needs to be implemented correctly. The most important condition for this to happen is that the results need to be considered good enough to detect skin cancer accurately. Since the binary logistic regression has shown a sensitivity of 86.1% at a specificity of 90.2%, this goal appears to be achieved. If this result can be combined with the integration of the binary logistic regression models that determine whether the lesion is an AK, BCC or an SCC, an implemented artifact covering the totality of these models can be used by doctors to support their decision making. A more elaborate description of opportunities of implementation is provided in section 7.3.

7.2 Limitations

The conclusions of this research project are bounded by the limitations of the project. A lot of the drawn conclusions are partly based on certain assumptions that are caused by some limitations. First of all, the biggest limitation in this research project is that the research is limited to the 6 categories of lesions that included the most images. A lot of other types of lesions exist that are not skin cancer, but have not been included in this research. Besides, melanoma have not been incorporated in this research, which would have a lot of added value, since melanoma is by far the most deadly type of skin cancer. [Linares et al. (2015)]

Secondly, not all the features that indicate skin cancer or that have an influence on the development of skin cancer, are included in the research. The research is limited to the features that can visibly be retrieved from the pictures, which does not give a complete overview. For example, features such as gender, medical record, sun exposure, age & lifestyle are not included, although literature shows those features have an impact on the probability of having skin cancer. [Van Der Geer et al. (2015)]

Moreover, as mentioned the presence or absence of certain features have been manually retrieved from the images. This is mainly due to time limitation, but it certainly causes errors in the estimation of the presence of certain features. Therefore, the assumption is made that the features are mostly correctly retrieved from the images and that the error rate of this estimation process

is comparable to the error rate that patients would have if they would name the features themselves.

Besides, an obvious limitation is the amount of images that have been used in this research. Although the amount of data has been enough to draw several suitable conclusions, additional images would probably have improved the results. The limited amount of data is mainly caused by the boundaries of the archive of MohsA and the limited amount of reliable images available on the internet. Access to a bigger database would probably improve results in the future. Moreover, more data on other lesions can improve the number of categories, which would also improve the results.

Finally, the amount of different locations on the body has been limited for the ease of this project. For example, no difference has been made between 'cheek' and 'chin'. These features are generalized to 'face' in order to avoid categories with too few samples. This generalization might have causes that certain features have been generalized into one categories while they have nothing in common for the prediction of skin cancer. For example, there is a major difference between a cheek and a lip in terms of cell structure, but they have both been generalized into the category 'face'. This limitation can be fixed in future research by the inclusion of more data as well.

7.3 Practical Implications

Separating between just two classes can already have a huge impact on planning processes at MohsA. If with 83% certainty can be stated whether a lesion is a non-melanoma skin cancer or not, this means a big step forward has been taken in the integration of machine learning in the dermatology. The developed model can be applied in two different ways.

Application 1: A mobile application for recognizing skin cancer from home

As described in the introduction, several mobile applications already exist for the recognition of skin cancer. For example, SkinVision is an application that is able to detect melanoma based on an uploaded picture on a smartphone [de Carvalho et al. (2019)]. The application achieves a 95% sensitivity rate, which is extremely high for a smartphone application [Udrea et al. (2020)]. Besides, for the detection of non-melanoma skin cancer, the application called 'OddSpot' is able to judge whether or not you have skin cancer based on several questions [Van Der Geer et al. (2015)]. A regression model conclusively gives a percentage on the probability of having actinic keratosis or basal cell carcinoma. However, a valuable application that has the ability to detect whether or not a non-melanoma skin cancer is present based on images is not yet available.

Therefore, this model could have a big impact to the currently existing mobile applications for the detection of skin-cancer. First of all, as currently known no publicly available application exists that is able to detect non-melanoma skin cancer based on regular image. Moreover, and this is where the added value of this particular research enters, the addition of techniques that improve the initial quality of the model can give a very nuanced view on the concerned lesion. The application could provide the user with information according to the following example structure.

The values are arbitrary and just to give an indication of how it might display:

Based on the image:
- The probability the lesion is **Skin Cancer** is **99%**
- The probability the lesion is **BCC** is **86%**

Based on the questions:
- The probability the lesion is **Skin Cancer** is **88%**
- The probability the lesion is **BCC** is **65%**

Advice:
- Visit a specialist within 3 weeks

This structure is just an example of how the information can be displayed. Other possibilities exist as well. The following structure shows a probability for each skin cancer type and for skin cancer overall. The prediction is based on the combination of the image and the questionnaire without displaying which part gave which indication:

- The probability the lesion is **Skin Cancer** is **92%**
- The probability the lesion is **AK** is **20%**
- The probability the lesion is **BCC** is **65%**
- The probability the lesion is **SCC** is **7%**

Advice:
- Visit a specialist within 3 weeks

All in all, multiple ways exist on what information should be displayed and how this information is corresponded to the target group in an effective way. However, it stands clear that several ways exist to integrate the results from this project into a mobile application.

Application 2: A support tool for expert dermatologists

Expert dermatologists that judge the presence of a certain type of non-melanoma skin cancer sometimes are not 100% sure whether the concerned lesion is a type of non-melanoma skin cancer or not. Then, this particular doctor can use the model as a confirmation tool when in doubt, by uploading the image and filling in the relevant questions. In this situation, the model does not necessarily need to be a mobile application. Therefore, better quality images can be used which is beneficial for the quality of judgement of the model.

7.4 Future Research

Several opportunities for future research exist. First of all, several features have not been added to the model, which has already been explained in section 7.2. The addition of these extra features

(gender, medical record, sun exposure, age, lifestyle, etc.), which already are widely accepted as predicting variables through literature [Van Der Geer et al. (2015)], could be very valuable to the performance and accuracy of the model. In order to do this, a major part of the time should be invested in data collection. These features are related to the habits and personality of the patients and can not be easily retrieved without the patients. Since the patients are mainly of a relatively older age, it might be hard to collect the data within a short amount of time.

Secondly, several machine learning techniques exist that have not been tested in this research. For example, an additional neural network can be used for the integration of additional features as well. Although such a network mostly functions as a black box, which make it hard to verify what features make a contribution, the prediction is that it would have a positive impact on the final classification, since convolutional neural networks are one of the most described machine learning techniques at the moment.

Another machine learning technique that can be tested in future research is generative adversarial networks (GAN). A generative adversarial network is a specific type of neural network that consists of two neural networks that compete against each other. This idea has been described first by Goodfellow et al. (2014). The first network is a generative network and the second network a discriminative network. The generative networks produces fake date based on the real data from the training set and the discriminative network evaluates those data and classifies is. The competitive factor between the two networks is that the generative network tries to produce data that increases the error rate of the discriminative network. In this way, both neural networks improve over time. The generative network becomes better at producing fake data, while the discriminative network improves by being confronted with those harder and harder data. The learning process is maintained through a constant loop of feedback between the networks through back-propagation.

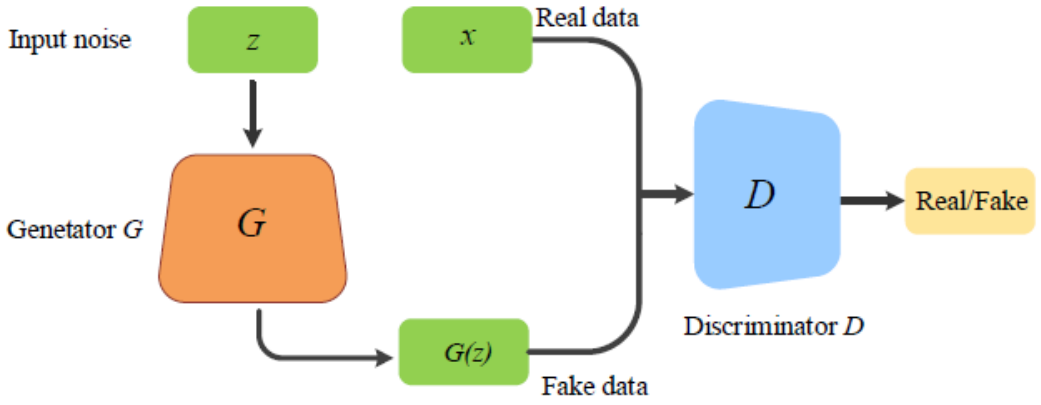

Figure 7.1: Architecture of GAN. [Zhu et al. (2018)]

A major advantage of this approach is that the network can be trained with a lot of extra data, because the generative neural network is producing additional data points. Since these data points are fake data points, the network is robust to over-fitting. A disadvantage of generative adversarial networks is that it's hard to balance because of the two networks. Moreover, the concept is relatively new so not a lot of research has been published in this field.

References

Agbai, O. N., Buster, K., Sanchez, M., Hernandez, C., Kundu, R. V., Chiu, M., Roberts, W. E., Draelos, Z. D., Bhushan, R., Taylor, S. C., et al. (2014). Skin cancer and photoprotection in people of color: a review and recommendations for physicians and the public. *Journal of the American Academy of Dermatology*, 70(4):748–762.

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., and Havel, J. (2013). Artificial neural networks in medical diagnosis.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210.

Brash, D. E., Ziegler, A., Jonason, A. S., Simon, J. A., Kunala, S., and Leffell, D. J. (1996). Sunlight and sunburn in human skin cancer: p53, apoptosis, and tumor promotion. *The journal of investigative dermatology. Symposium proceedings*, 1(2):136–142.

Brenner, M. and Hearing, V. J. (2008). The protective role of melanin against uv damage in human skin. *Photochemistry and photobiology*, 84(3):539–549.

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Fröhling, S., et al. (2019a). A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154.

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., et al. (2019b). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54.

Cracknell, M. J. and Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, 63:22–33.

de Carvalho, T. M., Noels, E., Wakkee, M., Udrea, A., and Nijsten, T. (2019). Development of smartphone apps for skin cancer risk assessment: progress and promise. *JMIR Dermatology*, 2(1):e13376.

Fernandez Figueras, M. (2017). From actinic keratosis to squamous cell carcinoma: pathophysiology revisited. *Journal of the European Academy of Dermatology and Venereology*, 31:5–7.

Flohil, S. C., De Vries, E., Neumann, M., Coebergh, J.-W., and Nijsten, T. (2011). Incidence, prevalence and future trends of primary basal cell carcinoma in the netherlands. *Acta dermato-venereologica*, 91(1):24–30.

Flohil, S. C., Van Der Leest, R. J., Dowlatshahi, E. A., Hofman, A., De Vries, E., and Nijsten, T. (2013). Prevalence of actinic keratosis and its risk factors in the general population: the rotterdam study. *Journal of Investigative Dermatology*, 133(8):1971–1978.

Furdova, A., Kapitanova, K., Kollarova, A., and Sekac, J. (2020). Periocular basal cell carcinoma-clinical perspectives. *Oncology Reviews*, 14(1).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Graupe, D. (2013). *Principles of artificial neural networks*, volume 7. World Scientific.

Guy Jr, G. P., Machlin, S. R., Ekwueme, D. U., and Yabroff, K. R. (2015). Prevalence and costs of skin cancer treatment in the us, 2002- 2006 and 2007- 2011. *American journal of preventive medicine*, 48(2):183–187.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842.

Hemanth, D. J., Anitha, J., Naaji, A., Geman, O., Popescu, D. E., et al. (2018). A modified deep convolutional neural network for abnormal brain image classification. *IEEE Access*, 7:4275–4283.

Hogarty, D. T., Su, J. C., Phan, K., Attia, M., Hossny, M., Nahavandi, S., Lenane, P., Moloney, F. J., and Yazdabadi, A. (2020). Artificial intelligence in dermatology—where we are and the way to the future: a review. *American journal of clinical dermatology*, 21(1):41–47.

Kallini, J. R., Hamed, N., and Khachemoune, A. (2015). Squamous cell carcinoma of the skin: epidemiology, classification, management, and novel trends. *International journal of dermatology*, 54(2):130–140.

Khatami, R., Mountrakis, G., and Stehman, S. V. (2016). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, 177:89–100.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Linares, M. A., Zakaria, A., and Nizran, P. (2015). Skin cancer. *Primary care*, 42(4):645–659.

Lisboa, I. N. D., Azevedo Macena, M. S. D., Conceicao Dias, M. I. F. D., Almeida Medeiros, A. B. D., Lima, C. F. D., and Carvalho Lira, A. L. B. D. (2016). Prevalent signs and symptoms in patients with skin cancer and nursing diagnoses. *Asian Pacific Journal of Cancer Prevention*, 17(7):3207–3211.

Lomas, A., Leonardi-Bee, J., and Bath-Hextall, F. (2012). A systematic review of worldwide incidence of nonmelanoma skin cancer. *British Journal of Dermatology*, 166(5):1069–1080.

Madan, V., Lear, J. T., and Szeimies, R.-M. (2010). Non-melanoma skin cancer. *The lancet*, 375(9715):673–685.

Marka, A., Carter, J. B., Toto, E., and Hassanpour, S. (2019). Automated detection of nonmelanoma skin cancer using digital images: a systematic review. *BMC medical imaging*, 19(1):21.

Marzuka, A. G. and Book, S. E. (2015). Basal cell carcinoma: pathogenesis, epidemiology, clinical features, diagnosis, histopathology, and management. *The Yale journal of biology and medicine*, 88(2):167–179.

Maxwell, A. E., Warner, T. A., and Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9):2784–2817.

Montantes, J. (2020). 3 reasons to use random forest over a neural network–comparing machine learning versus deep learning.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Shoimer, I., Rosen, N., and Muhn, C. (2010). Current management of actinic keratoses. *Skin Therapy Lett*, 15(5):5–7.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tietjen, G. E., Peterlin, B. L., Brandes, J. L., Hafeez, F., Hutchinson, S., Martin, V. T., Dafer, R. M., Aurora, S. K., Stein, M. R., Herial, N. A., et al. (2007). Depression and anxiety: effect on the migraine–obesity relationship. *Headache: The Journal of Head and Face Pain*, 47(6):866–875.

Udrea, A., Mitra, G., Costea, D., Noels, E., Wakkee, M., Siegel, D., de Carvalho, T., and Nijsten, T. (2020). Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *Journal of the European Academy of Dermatology and Venereology*, 34(3):648–655.

Van Der Geer, S., Kleingeld, P. A. M., Snijders, C. C., Rinkens, F. J., Jansen, G. A., Neumann, H. M., and Krekels, G. A. (2015). Development of a non-melanoma skin cancer detection model. *Dermatology*, 230(2):161–169.

Vandiver, A. R., Irizarry, R. A., Hansen, K. D., Garza, L. A., Runarsson, A., Li, X., Chien, A. L., Wang, T. S., Leung, S. G., Kang, S., et al. (2015). Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome biology*, 16(1):1–15.

Vimercati, L., De Maria, L., Caputi, A., Cannone, E. S. S., Mansi, F., Cavone, D., Romita, P., Argenziano, G., Di Stefani, A., Parodi, A., et al. (2020). Non-melanoma skin cancer in outdoor workers: A study on actinic keratosis in italian navy personnel. *International Journal of Environmental Research and Public Health*, 17(7):2321.

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Springer-Verlag London, UK.

Wu, Y., Jiang, X., Wang, S., Jiang, W., Li, P., and Ohno-Machado, L. (2015). Grid multi-category response logistic models. *BMC medical informatics and decision making*, 15(1):10.

Yaldiz, M. (2019). Prevalence of actinic keratosis in patients attending the dermatology outpatient clinic. *Medicine*, 98(28).

Yang, J.-J., Li, J., Shen, R., Zeng, Y., He, J., Bi, J., Li, Y., Zhang, Q., Peng, L., and Wang, Q. (2016). Exploiting ensemble learning for automatic cataract detection and grading. *Computer methods and programs in biomedicine*, 124:45–57.

Zhang, Q., Yang, L. T., Chen, Z., and Li, P. (2018). A survey on deep learning for big data. *Information Fusion*, 42:146–157.

Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232.

Zhu, L., Chen, Y., Ghamisi, P., and Benediktsson, J. A. (2018). Generative adversarial networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9):5046–5063.

A Decision Tree