DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2018

Evaluating a deep convolutional neural network for classification of skin cancer

JOAKIM BOMAN

ALEXANDER VOLMINGER

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Evaluating a deep convolutional neural network for classification of skin cancer

JOAKIM BOMAN AND ALEXANDER VOLMINGER

Bachelor in Computer Science
Date: June 6, 2018
Supervisor: Pawel Herman
Examiner: Örjan Ekeberg
Swedish title: Utvärdering av ett deep convolutional neural network för klassificering av hudcancer
School of Electrical Engineering and Computer Science

Abstract

Computer-aided diagnosis (CAD) has become an important part of the medical field. Skin cancer is a common and deadly disease that a CAD system could potentially detect. It is clearly visible on the skin and therefore only images of skin lesions could be used in order to provide a diagnosis. In 2017, a research group at Stanford University developed a deep convolutional neural network (CNN) that performed better than dermatologists during classification of skin lesions.

This thesis makes an attempt at implementing the method provided in the Stanford report and evaluate the performance of the CNN during classification of skin lesion comparisons not tested in their study. The previously unseen binary classification use cases are melanoma versus solar lentigo and melanoma versus seborrheic keratosis. Using transfer learning, Inception v3 was trained for various skin lesions. The CNN was trained with 16 training classes. During validation of the CNN, an accuracy of 68.3% was achieved during a 3-way classification. Testing the same comparisons as the Stanford study an accuracy of 71% was achieved for melanoma versus nevus and 91% for seborrheic keratosis versus basal and squamous cell carcinoma. The accuracy results for the new comparisons were 84% for seborrheic keratosis versus melanoma and 83% for solar lentigo versus melanoma.

The results suggest that out of the binary classifications performed in this study, nevus versus melanoma is the most difficult for the CNN. It should be noted that our results were different from the Stanford study and that more statistical methods should have been used when obtaining the results.

Sammanfattning

Computer-aided diagnosis (CAD) har blivit en viktigt del av det medicinska området. Hudcancer är en vanlig och dödlig sjukdom som ett CAD system potentiellt kan upptäcka. Den är klart synlig på huden och därför skulle endast bilder av hudskador kunna användas för att ge en diagnos. År 2017 utvecklade en forskningsgrupp från Stanford University ett deep convolutional neural network (CNN) som presterade bättre än dermatologer vid klassificering av hudskador.

Detta kandidatexamensarbete gör ett försök till att implementera metoden tillhandahållen i Stanford rapporten och utvärdera CNN:ets resultat vid klassifikation av hudskador som inte testades i deras studie. De binära fall som tidigare inte har testas är melanoma emot solar lentigo och melanoma emot seborrheic keratosis. Med hjälp av transfer learning tränades Inception v3 för olika hudskador. CNN:et tränades med 16 typer av hudförändringar. I valideringsprocessen uppmättes en korrekthet på 68.3% under 3-vals klassifikation. I tester av samma typ av jämförelser som i Stanford studien uppmätes en korrekthet på 71% för melanoma emot nevus och 91% för seborrheic keratosis emot basal and squamous cell carcinoma. Resultatet av de nya jämförelserna var 84% för seborrheic keratosis emot melanoma och 83% för solar lentigo emot melanoma.

Resultaten tyder på att av de binära klassificeringarna utförda i denna studie, är nevus emot melanoma den svårast för CNN:et. Det bör noteras att våra resultat skilde sig från Stanford studien och att mer statistiska metoder borde använts för framtagningen av resultaten.

Contents

Chapter 1

Introduction

Today, computer-aided decision systems have become important when evaluating and diagnosing medical images [1]. For example, computer-aided diagnosis (CAD) is part of the routine when detecting breast cancer on mammograms in the United States [2]. CAD is also one of the major research subjects in medical imaging and diagnostic radiology [2]. An accurate CAD system can be used for early detection of a disease and thereby allow for earlier treatment, which could save lives [3]. For example, the ability to effectively treat and cure cancer is directly dependent on the ability to detect cancers at their earliest stages [4].

Cancer is a collection of related diseases where diagnosis and treatment are of great interest due its widespread occurrence [5]. In 2012, there were 14 million new cases of cancer and 8.2 million cancer-related deaths worldwide. This makes cancer one of the most common causes of death for humans [6]. Skin cancer is the most common type of cancer and usually forms in skin that has been exposed to sunlight, however it can occur on any part of the body [7]. Skin cancer begins in the epidermis (outer layer of the skin) and is therefore clearly visible. This means that a CAD has potential to use only images of the skin lesion, without any other information, in order to give a preliminary diagnosis.

Recently, a study performed at Stanford University [8] developed a deep convolutional neural network (CNN) that performed better than dermatologists when classifying keratinocyte carcinomas versus benign seborrheic keratoses and malignant melanomas versus benign nevi. However, it is still unknown how the CNN performs during classification of other skin diseases.

1.1 Problem statement

This study will investigate the classification accuracy of a CNN developed using the method presented by [8]. The CNN will be designed to distinguish melanoma from seborrheic keratosis and solar lentigo, which can be difficult [9, 10]. By doing so, this paper will establish new test results for the method proposed by the Stanford study [8]. This is important since the algorithm needs to be able to distinguish between multiple different benign and malignant lesions in order to provide a correct medical diagnosis. Therefore, we intend to study how the performance of a state-of-the-art CNN depends on various types of skin lesions.

1.2 Scope and objectives

Only open-access datasets will be used to train, validate and test the CNN. In order to set up a similar environment to [8], the same images and are used when possible. However, not all datasets used by [8] were available and some new datasets are used in order to increase the number of images for the diseases investigated in this report.

Performance of the CNN will be tested on diseases that have been acknowledged by [9, 10] to be hard to distinguish from melanoma. The focus of this thesis is to compare the accuracy of the CNN when classifying different skin lesions.

There are two main objectives, create the skin lesion dataset and use transfer learning on Google's Inception v3. The skin lesion dataset will be preprocessed according to the same method used by [8]. Because of the relative small amount of skin lesion pictures that are available, transfer learning will be used. Inception v3 is the network used for training since [8] got the best results with this network.

1.3 Thesis outline

In the next chapter, background information and previous work relevant for this thesis is presented. The third chapter describes the procedure of this study. It contains information about the datasets used and how the CNN was trained, validated and tested. The fourth chapter presents the results gathered from the experiment. In the fifth chapter the results are analysed and limitations, ethics and sustainability are discussed. Chapter six presents the conclusion and suggests further research.

Chapter 2

Background

2.1 Skin cancer

Cancer refers to a collection of related diseases where some of the body's cells begin to divide without stopping and spread into surrounding tissues [11]. Skin cancer is the most common type of cancer and can be highly malignant [12]. It is most often caused by ultraviolet radiation from the sun, which damages the DNA in skin cells. The damaged DNA then triggers mutations that makes the skin cells multiply and form tumors. Skin cancer can also occur from genetic defects [13].

Various classifications of skin cancer exist. Some examples are melanoma, basal and squamous cell carcinoma out of which melanoma is the most deadly [14]. Basal and squamous cell carcinoma rarely spread beyond the original tumor site [7]. Melanoma represent solely 4% of all skin cancers, however it is chargeable for 75% of all skin cancer deaths [14]. Melanoma is more aggressive compared to the other skin cancer types and likely spreads to nearby tissues [7]. Early detection of melanoma is critical, as the estimated 5-year survival rate drops from over 99% if detected during earliest stages to about 14% if detected in its latest stages [8].

Today, skin cancer diagnoses are mainly determined visually. First, an initial clinical screening is performed which potentially is followed by dermoscopic analysis, a biopsy and a histopathological examination [8].

2.1.1 Skin lesion classification

In skin lesion classification, features play an important role [15]. Many different features can be taken into account in the context of skin. Some of the categories are color features, contour features, dermal features, geometric features, histogram features, texture features and the ABCD rule features which analyses the asymmetry, border irregularity, color variation and diameter [15].

2.1.2 Clinical terms and tests

In order to evaluate a clinical test the terms sensitivity and specificity can be used. The sensitivity of a clinical test refers to the ability of the test to correctly identify the patients with the disease [16].

$$sensitivity = \frac{true\ positive}{positive}$$

The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease [16].

$$specificity = \frac{true\ negative}{negative}$$

The four terms used to describe the comparison metrics sensitivity and specificity are described as follow [8]:

1. *true positive*: the number of correctly predicted malignant lesions

2. *positive*: the number of malignant lesions shown

3. *true negative*: the number of correctly predicted benign lesions

4. *negative*: the number of benign lesions shown

2.2 Technical background

2.2.1 Neural networks

Artificial Neural Networks (ANN) is an attempt to mimic the neurons of the brain. However, the models used have many simplifications and thus they do not reflect the true behaviour of the brain. Development of ANN had its first peak in the 1940s and development has since had its ups and downs [17].

By using the mathematical model:

$$y=\theta\left(\sum_{j=1}^{n} w_j x_j - u\right),$$

the weighted sum of signals from other connected nodes is computed and a model of a neuron is created. The nodes are connected in two main patterns. In feed-forward networks no loops occur and in recurrent networks loops occur. There is also a multilayer feed-forward network where there is an initial input stage, hidden layers and an output layer.

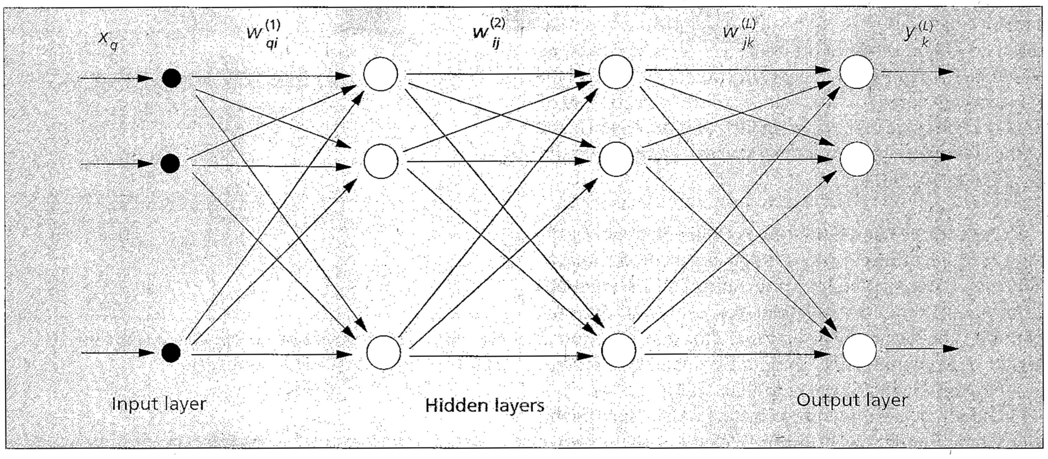

Figure 1 - A feed-forward network with two hidden layers [17].

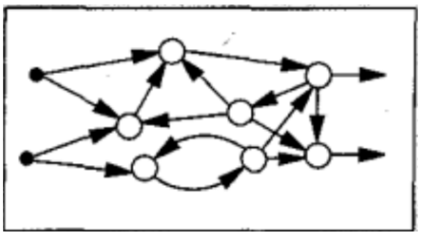

Figure 2 - An example of a recurrent network [17].

The learning part of ANN comes in when finding what weight each neuron should have in order to get the sought result. In the learning process ANNs tend to find patterns in the supplied examples. One way of doing learning with an ANN is supervised learning. In supervised learning the network is supplied with labeled data for all possible input. The weights are then calculated to produce answers according to the labeled data. If the examples are too few, over-fitting may occur for the network. This is when the ANN work well for the training data, but poorly on unseen examples. There are 4 main learning rules, error-correction is one of them. Error-correction falls under supervised learning and the basic idea is to use the error to update the weights and lower the error over each iteration. An algorithm based on this is backpropagation [17]. In this algorithm the errors get propagated through the layers backwards. The algorithm is based on gradient decent to minimize the error [17].

2.2.2 Deep neural networks

A deep neural network is a multi-layer neural network with many hidden layers. The idea of DNN has been around for many years, but it is only recently that many of the problems associated with the technique have been solved. This is mainly because of new learning algorithms and an increase of compute power. The three main problems for multilayer networks were: vanishing gradient, overfitting and computational load [18]. The simplest way to improve a deep neural network (DNN) is by increasing its size in both depth and width, but this simple solution comes with two drawbacks. Increased size will make the networks more prone to overfitting and it will also greatly increase the computer power required [19].

2.2.3 Convolutional neural networks

A convolutional neural network (CNN) tries to imitate how the visual cortex of the brain recognize images. To get better results with image classification image, feature extraction should be used [18]. Before CNNs existed, these feature extractors were designed by experts in each field of the images to be classified. However, with CNNs the feature extractor is included in the training process. The feature extractors consist of several convolutional layers and pooling layers. The convolutional layer can be seen as a digital filter. The pooling layer reduces the dimension of the image by combining neighbouring pixels into a single pixel [18]. CNN is one of the main reasons why in the last couple of years there have been major advances in image recognition. LeNet5 set what now have become standard structure for CNN [20]. The structure has stacked convolutional layers, which can be followed by contrast normalization and max-pooling layers. These are then followed by fully-connected layer(s).

Compared to feed-forward network with similarly sized layers a CNN has fewer parameters and connections. This makes them easier to train, but theoretically their best performance is somewhat less than a feed-forward network [20]. CNNs are computational heavy when applied on a large scale over high resolution images, but with the GPUs since 2012 and optimized versions of 2D convolution it is possible to do this with reasonable computational resources [20].

2.2.4 Transfer learning

Stuart J. Russell, a Professor of Computer Science at University of California, Berkeley once said [21, as cited in]:

> It is easier to learn to play chess already knowing how to play checkers, or to learn Spanish already knowing Italian.

This is something you can take advantage of by using transfer learning. CNNs can either be trained from scratch where all its parameters are tuned for the problem, or they can be tuned towards the problem from an already pre-trained CNN. Transfer learning is often used with CNN in the way that all layers are kept except the last one, which is trained for the specific problem. This method can be particularly useful for medical applications since it does not require as much training data, which can be hard to get in medical situations [22].

Both [23] and [24] are examples of when there is better performance with transfer learning compared to learning a CNN with a small dataset.

2.2.5 Inception Architecture

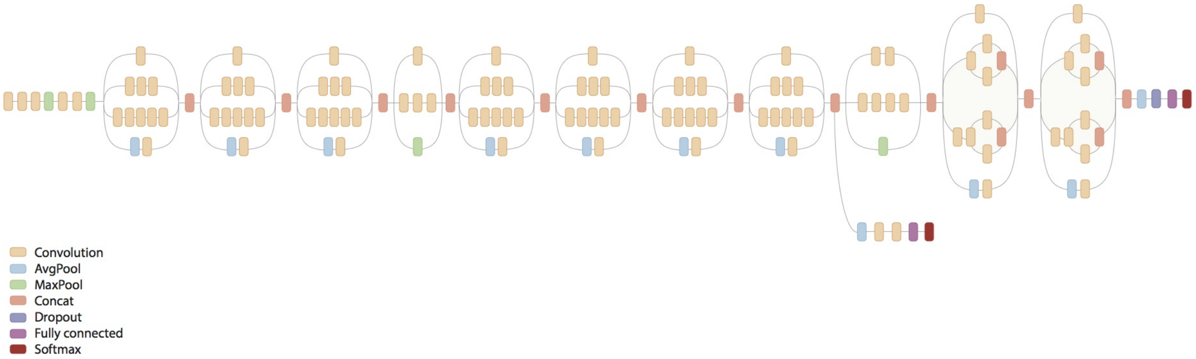

Figure 3 - Inception v3 architecture [25].

GoogleNet Inception v3 is a CNN object renegotiation architecture that is pre-trained with a dataset with around 1,28 million pictures [26].

Inception is a codename for an efficient deep neural network architecture for computer vision. Here a “Inception module” is introduced, which is a new level of organization. More depth in the network is also added [19]. The Inception architecture have been found to perform very good even with low computational power [27].

Conventional convolutional layers use linear functions and are thus not able to learn non-linear functions. In [28] it was proposed a way a way to make convolutional filters more complex. With these contributions it was possible for the filters to learn non-linear functions. The results in [28] also made CNNs less prone to over-fitting. Building on these results [28] the first Inception version-1 architecture was developed. The main concept of the architecture is finding a approximation of the optimal structure in a CNN and how to build this by already available components [19]. Building further on the idea of the Inception Architecture version 2 and 3 was introduced. Both yielding better results than the original [26]. There is also a version 4 of the Inception architecture where residual connections are combined with the architecture. The architecture is also streamlined, both helps reduce the computational power needed in the training process. Version 4 also achieves better results than the previous versions [27].

2.3 Related work

In a study performed at Stanford University [8] the authors retrained Inception v3 using transfer learning with labeled pictures of skin lesions. The performance of the CNN was then compared to 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases [8]. The deep learning CNN outperformed the average dermatologist at skin cancer classification using photographic and dermoscopic images. For example, when the algorithm was validated using a three-class disease partition, their CNN achieved 72.1 ± 0.9% overall accuracy, while two dermatologists attained 65.56% and 66.0% accuracy on a subset of the validation set [8]. When using a nine-way classification, their CNN achieved an accuracy of 55.4 ± 1.7%, while the dermatologists attained lower accuracies at 53.3% and 55.0%. The study also performed a couple of binary comparisons between diseases. The results were presented in ROC-curves, two of the graphs are shown in Figure 4.

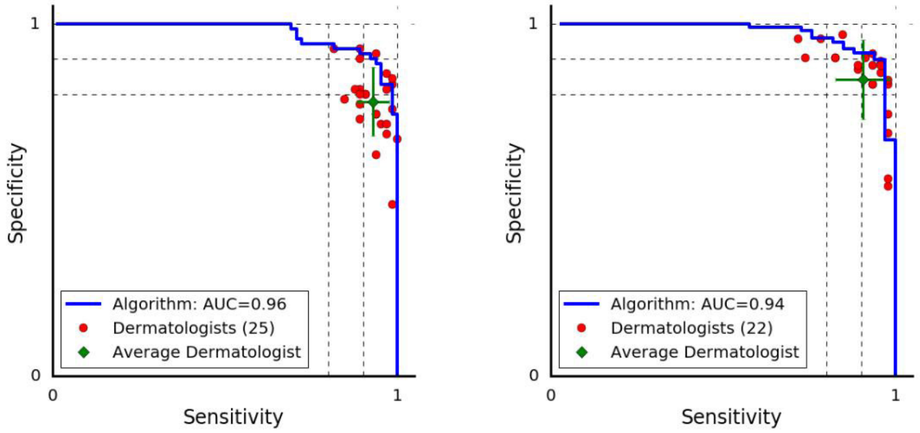

Figure 4 - ROC-curves from the results acquired by [8]. The left graph is the ROC curve for their test set with 65 carcinomas and 70 seborrheic keratosis images. The right graph is the ROC curve for their test set with 33 melanoma and 97 nevus images.

A similar conclusion was presented in a study by Nylund [29]. In the study Nylund used a common convolutional neural network for image classification called AlexNet [29]. The overall accuracy was 55.0% for a 23-way classification, which is close the result of the 9-way classification done by [8]. Nylund concluded that the performance of convolutional neural networks are comparable to practicing dermatologists and state of the art machine learning methods in binary classification accuracy, benign – melanoma, with only little pre-processing and tuning [29].

In the study by Ridell and Spett [30], a CNN was trained based on Google Inception v3 in order to detect skin cancer. It was then investigated how the accuracy of classifying between benign nevus and malignant melanoma is affected by the size of the training dataset. Multiple image sizes were tested ranging from 200 to 1600 number images which resulted in accuracy between 70.8% and 77.5%. The conclusion was that the accuracy seemed to be higher with a larger dataset [30].

Chapter 3

Method

The method outlined in this section is inspired by the method described in Stanford study [8] and when possible the same steps have been taken.

3.1 Dataset

The images used in this study comes from multiple different open-access sources. From the ISIC Dermoscopic Archive [31] 23,647 images were received. The ISIC dataset contains images labeled with the diagnosis of the skin lesion as well as other information about the image. Images labeled as ″other″ and ″unknown″ were not used since the skin lesions in these groups could be of any disease.

An additional 16,826 images were downloaded from DermQuest [32] and 4,336 images from the Dermatology Atlas [33]. Furthermore, a total of 1,948 images were taken from the websites DermaAmin [34], Dermoscopy Atlas [35], Global Skin Atlas [36], Hellenic Dermatological Atlas [37], Medscape [38], Regional Derm [39], Skinsight [40] and the pH2 database [41].

Images were put into folders depending on their diagnosis which was retrieved from image metadata. The folders were structured in the same way as the taxonomy found on the Github repository [42], which belongs to one of the authors of [8], Sebastian Thrun. The taxonomy scheme was derived from dermatologists that arranged the skin lesions based on clinical and visual similarity. It is arranged in a tree structure where the three root nodes are benign lesions, malignant lesions and non-neoplastic lesions. The children of the root nodes are the individual diseases.

The taxonomy was then leveraged to partition the individual diseases into training classes. This generated training classes with diseases that are clinically and visually similar since all training classes are descendants of the root nodes. Each training class was ensured to have a minimum of 200 images so that the CNN would have sufficient training data for each class. If a skin lesion class had less than 200 images it was grouped together with other skin lesions close to it in the taxonomy. A maximum class size of 1000 was used for the training classes. By using this partition a balance is achieved: training classes are not too small for the CNN to learn them properly and classes are not too coarse, which could make the CNN biased towards a specific class. The partitioning resulted in 16 training classes. Images were randomly chosen to either be in the training or validation set, with a 10% chance of being put into the validation set. Blurry and and far-away images were removed from the validation and test set. However, they were still used in training. When selecting images for the test set, images from the ISIC Dermoscopic Archive [31] were selected when possible since they were deemed to have the best quality. Most of the images in the test set have a biopsy proven diagnosis. The number of images used in the test set for each skin lesion type is shown in Table 1.

Skin lesion	**No. images**
Nevus	ISIC: 97
Seborrheic keratosis	ISIC: 19 DermQuest: 69 Skinsight: 1 Regional Derm: 6 Dermoscopy Atlas: 2
Solar lentigo	ISIC: 20
Melanoma	ISIC: 33
Basal cell carcinoma	ISIC: 35
Squamous cell carcinoma	ISIC: 30

Table 1 - Number of images used in the test set.

3.2 Setting up the CNN and training process

The final layer of Inception v3, see Figure 3, was retrained with the skin lesion data set. Transfer learning was used because of the relative small amount of data available. The CNN was trained with backpropagation. All layers were set to use the same global learning rate of 0.001 and a decay factor of 16 every 30 epochs. RMSProp was used with a decay 0.9, momentum 0.9 and epsilon 0.1. Batch size was set to 100.

Google's TensorFlow [43] was used to train, validate and test the CNN. The images were augmented by randomly rotating between 0 and 359 degrees during training. Furthermore, for each image the largest inscribed rectangle, see Figure 5, was cropped from the image and flipped vertically with a probability of 0.5.

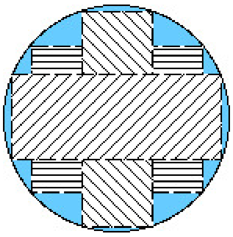

Figure 5 - Variations of the largest inscribed rectangle [44].

All images were resized to 299x299 pixels since this is the original dimension Inception v3 was originally trained on.

3.3 Validation process

In order to compare the performance of the CNN to [8], the accuracy of the validation data set was measured with both a 16-way classification and a 3-way classification. The 3-way classification was based on the first level of the taxonomy [8]. The 16-way classification was received from doing the partitioning on the taxonomy. Below are tables showing the disease classes in the two classification strategies:

1. Benign skin lesions
2. Malignant skin lesions
3. Non-neoplastic skin lesions

Table 2 - Disease classes for the 3-way classification [8].

1. Acne Folliculitis Hidradenitis And Diseases Of Appendegeal Structures-bucket
2. Acne Vulgaris
3. Basal Cell Carcinoma
4. Dermal Tumor Benign
5. Eczema Spongiotic Dermatitis
6. Epidermal Tumor Benign-bucket
7. Epidermal Tumor Malignant-bucket
8. Inflammatory-bucket
9. Melanoma
10. Nevus
11. Pigmented Lesion Benign-bucket
12. Pigmented Lesion Malignant-bucket
13. Psoriasis Pityriasis Rubra Pilaris And Papulosquamous Disorders
14. Rosacea Rhinophyma And Variants
15. Seborrheic Keratosis
16. Seborrheic Keratosis-bucket

Table 3 - The 16 disease classes received from doing the partitioning described in section 3.1. The bucket classes are combination of multiple similar skin lesions in order to get the correct number of images for the training class.

The 16-way classification accuracy was calculated each epoch. The network receiving the highest accuracy compared to previous validation accuracies was saved. After training, 3-way classification accuracy was calculated on the validation data set. A 3-way accuracy for an image was obtained by calculating the probability for each root node. If the root node with the largest probability contains the label of the image according to the taxonomy, this was seen as a successful classification. To obtain the network validation accuracy, the number of successful labeled images was divided with the total number of images in the validation set. The same data augmentation as in the training process was also done to each validation image.

3.4 Test process

The clinical test in this study consists of the CNN receiving images as input and returning a result in form of a label, either non-malignant or malignant. The test is binary in order to simulate the process dermatologists have when choosing whether or not a biopsy should be performed [8]. A malignant score was calculated by summarizing the probabilities of the classes that contained images of malignant skin lesions. If the malignant score was greater than 50% the image was labeled malignant. A relative operating characteristic (ROC) curve was used to measure the association between observed and diagnosed presence of each characteristic. The ROC curve was created by plotting the sensitivity against the specificity sweeping over various thresholds between 0 to 1. In addition to the curve we also get a metric called area under the curve (AUC). An AUC of 1 means that the classifier can differentiate perfectly between diseased and non-diseased. If the AUC is 0.5, the classifier cannot see any difference between the two distributions [45]. In the ROC curves a dotted blue line will be used to show where the area is less than 0.5. For each image in the test set data augmentation was done in the same way as for the training and validation sets.

A total of four binary comparisons were made and can be found in Table 4. Comparisons 1 and 2 were tested in order to compare the performance of the CNN in this study to the CNN presented by [8]. Comparisons 3 and 4 were performed since both seborrheic keratosis and solar lentigo are known to be difficult to distinguish from melanoma [9, 10].

Comparison	Skin lesion 1	Skin lesion 2
1	Nevus	Melanoma
2	Seborrheic keratosis	Basal and squamous cell carcinoma
3	Seborrheic keratosis	Melanoma
4	Solar lentigo	Melanoma

Table 4 - The four binary comparisons performed in this study.

Chapter 4

Results

4.1 Validation results

After training, the CNN was tested on the validation set with 1390 images. Two tests were performed. In the first one, a 16-way classification was done, where the CNN could classify between all the training classes. In the second test, the different training classes were combined to three classes, as described in section 3.3. Note that the CNN has more options to choose from when doing the 16-way classification. This means that it is less likely for the CNN to guess the correct classification. Therefore the accuracy results cannot be directly compared. The validation accuracy for the CNN when using a 3-way classification respectively a 16-way classification is shown in Table 5. The CNN achieved a higher accuracy when doing the 3-way classification.

Classification	Accuracy
3-way	68.3%
16-way	52.0%

Table 5 - Validation accuracy for the CNN. In the 3-way classification the CNN has 3 classes to choose from and in the 16-way classification there are 16 classes.

4.2 Binary classification accuracy

In Figure 6 the ROC curves for each binary comparison are displayed in separate windows. The accuracy of each binary classification, which is measured by the area under the ROC curve, is show in decimal form in Table 6. The CNN did not achieve a lower AUC than for nevus versus melanoma, which is represented by the ROC curve in Figure 6.a. No higher AUC was measured than for seborrheic keratosis versus basal and squamous cell carcinoma, which is represented by the ROC curve in Figure 6.b. For the two comparisons seborrheic keratosis versus melanoma and solar lentigo versus melanoma, which can be found in Figure 6.c and 6.d, the AUC was similar.

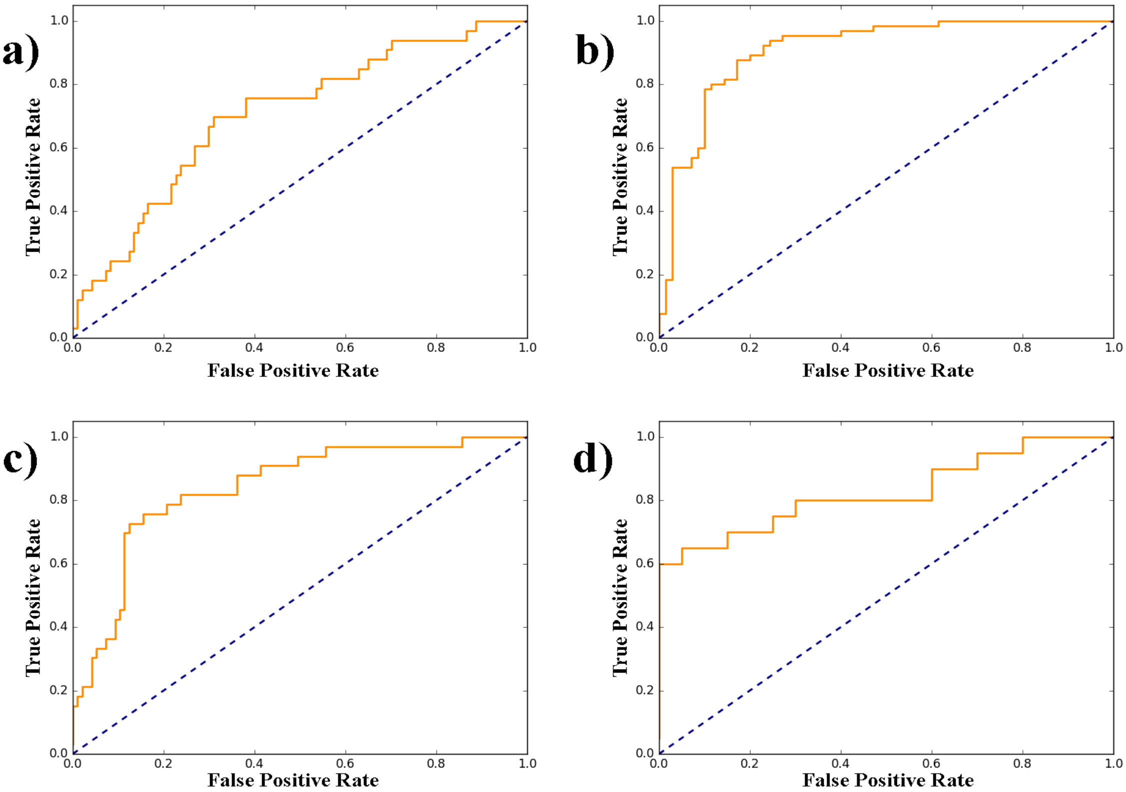

Figure 6 - ROC curves for the CNN tested on four different binary comparisons. Each window a to d represents a comparison. The comparisons and number of images used for testing are as follows:

a) 97 pictures of nevus and 33 pictures of melanoma.

b) 70 pictures of seborrheic keratosis and 65 pictures of basal and squamous cell carcinoma.

c) 97 pictures of seborrheic keratosis and 33 pictures of melanoma.

d) 20 pictures of solar lentigo and 20 pictures of melanoma.

Comparison	AUC
Nevus versus melanoma	0.71
Seborrheic keratosis versus basal and squamous cell carcinoma	0.91
Seborrheic keratosis versus melanoma	0.84
Solar lentigo versus melanoma	0.83

Table 6 - AUC for the ROC curves in Figure 6.

As seen in the figure and table above, the CNN seems to achieve a higher AUC when classifying melanoma against solar lentigo and seborrheic keratosis when compared to nevus. For solar lentigo the CNN outperformed nevus with 12 percentage points and for seborrheic keratosis the accuracy was 13 percentage points higher.

Chapter 5

Discussion

5.1 Key findings

The results indicate that the CNN obtained by this study does not find it harder to classify malignant melanoma against other benign lesions than nevus. As stated by [9, 10], solar lentigo and seborrheic keratosis are visually similar to melanoma. Therefore, the results strengthen the credibility of the CNN during classification of skin lesions. However, none of the comparisons to melanoma achieved an accuracy better than the test with seborrheic keratosis versus basal and squamous cell carcinoma. This indicates that it is more difficult for the CNN to separate benign lesions from melanoma than it is to separate them from the carcinoma cancer types.

5.2 Comparison to related work

The intent was that the CNN in this study would classify the binary comparisons nevus versus melanoma and seborrheic keratosis versus basal and squamous cell carcinoma with an accuracy similar to the CNN presented by [8]. This was not the case, the CNN accuracy for nevus versus melanoma was 23 percentage points worse and seborrheic keratosis versus basal and squamous cell carcinoma was 5 percentage points worse. However, the accuracy for nevus versus melanoma is comparable to the results of [30] who got an accuracy of 74.3 ± 1.3% when training on 1000 images, which is the same size as the maximum size used for our training classes. The CNN by [30] was only trained on nevus and melanoma, which could explain why they got a slightly higher accuracy in their binary comparison.

The validation results show that the CNN has slight lower validation accuracy of 52.0% during 16-way classification when compared to the similar 9-way classification by [8] and the 23-way classification by [29]. This is also slightly lower than the 9-way accuracy achieved by the two dermatologists in [8]. It is important to note that the 16-way classification gives the CNN more options to choose from during classification, so a lower classification accuracy can be expected when compared to 9-way. These accuracy results are therefore not completely comparable and it is possible that the CNN in this study would perform better during 9-way classification.

During 3-way classification the CNN got an accuracy of 68.3% which is lower than the accuracy acquired by [8]. However, our network performed better than the two dermatologists in [8]. The validation results thereby show that the performance of the CNN in this study is similar to the CNN and dermatologists in [8] when classifying skin lesions.

5.3 Limitations

During training the images used were all from open-access dermatology repositories. However, not all images used by [8] were available for this thesis. For example the Edinburgh Dermofit Library and data from the Stanford Hospital were not openly available. This means that the CNN could not be tested on the same test set. It was also not possible to obtain the same number of training classes. This is because of a lack of skin lesion images when compared to [8]. There was also a lack of time for this study which meant the it was not made sure that images of the same lesion with different angles were split between the training and validation sets.

A lack of compute power meant the augmentation was only done one time during training. Doing augmentation more often could mean that the performance of the CNN would increase. It was not clear how [8] carried out their augmentation, which meant that we choose the approach with only one augmentation. In the augmentation there also is some randomness involved, which means that the network performance can slightly vary between different instances.

When obtaining our results more statistical tests should have been done. This is a serious flaw in our report that diminish the conclusions that can be drawn from the results. Cross-validation should have been used to obtain the 3-way and 16-way accuracy for the CNN. This was done by the Stanford study [8]. Due to the time constraints with this study it was not performed. Something similar to cross-validation should also have been used when obtaining our ROC curves. This should have been done to get more consistent results on the ROC curves. Therefore our results lack statistical evidence. Having this would have allowed for more reasoning about the results and showed statistical significant differences between the AUC for the different ROC curves. It appears the Stanford study [8] did not too use any cross-validation when obtaining their ROC curve. They did however run a second test for each binary comparison with a larger test set, getting a second ROC curve for each test.

The method used for training, validating and testing the CNN mimicked the method used by [8] as well as possible. This means that most of the configuration is the same but there were multiple details missing in their method.

Because of the variance in the dataset, method and augmentation the environment will differ and may have implication on the performance, which means that there cannot be a solid comparison of our results against [8].

5.4 Ethics and sustainability

The skin lesion pictures used in this study are all from real persons. Even though they have given their consent to be in the datasets, one should take great care when using the images. Especially with the images showing more of the person than the skin lesion itself. One should aim to keep the subjects' integrity. Another aspect of the pictures in the dataset is that most were of skin with light color. When other skin colors are presented to the CNN its performance may be effected. The variation of performance could create a discrimination between people of different skin colors.

It is also important to question where the responsibility lies during the classification decision. This could be a potential life or death decision and therefore it can be hard to decide who is to blame when a skin lesion is wrongly classified. It could potentially be the CNN, the creators or someone else. Even dermatologists make mistakes during classification. If the CNN makes fewer mistakes than dermatologists maybe that redeems the life or death responsibility potentially given to the CNN.

The potential of using the CNN in this study is that everyone with a camera and computer can obtain the classification service traditionally provided by dermatologists. This would make the care more accessible, as you are no longer required to travel to the hospital in order to get a preliminary diagnosis. Furthermore, you would no longer need fully trained dermatologists to classify skin lesions, which allows the dermatologists to use their time for other tasks.

Chapter 6

Conclusion

The results indicate that a CNN developed by the method presented in [8] would not perform worse for binary classification of solar lentigo versus melanoma and seborrheic keratosis versus melanoma, compared to nevus versus melanoma. Comparing the new binary classifications to seborrheic keratosis versus basal and squamous cell carcinoma the CNN would perform slightly worse. Therefore, nevus versus melanoma seems to be the hardest classification for the CNN out of the ones tested. However, this is not certain since the study was not able to mimic the method by [8] in every detail. A lack of statistical evidence in the results also diminish the conclusion, since no statistical significant differences between the AUCs can be established.

6.1 Future research

There is an opportunity of continued study in trying to achieve a CNN with greater or equal accuracy compared to [8] during classification of nevus versus melanoma. Thereafter do the same binary comparisons as presented in this report. It would also be interesting to compare the performance of dermatologists to our results for classification of solar lentigo and seborrheic keratosis versus melanoma. Furthermore, we were only able to find two skin lesions that were confirmed to be visually similar to melanoma. Dermatologists may be able to find other binary comparisons that would need testing before using the CNN in a real clinical setting.

Other research that could be done is investigating how the CNN performs for skin of different color. This would be important to do in order to see if the CNN could be used by all humans.

Bibliography

[1] Masood A, Ali Al-Jumaily A. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. International journal of biomedical imaging. 2013;2013.

[2] Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics. 2007;31(4-5):198–211.

[3] Abdel-Zaher AM, Eldeib AM. Breast cancer classification using deep belief networks. Expert Systems with Applications. 2016;46:139–144.

[4] Wulfkuhle JD, Liotta LA, Petricoin EF. Early detection: proteomic applications for the early detection of cancer. Nature reviews cancer. 2003;3(4):267.

[5] Choi YE, Kwak JW, Park JW. Nanotechnology for early cancer detection. Sensors. 2010;10(1):428–455.

[6] National Cancer Institute. Cancer Statistics; 2017. Accessed: 2018-04-22. Available from: `https://www.cancer.gov/about-cancer/understanding/statistics`.

[7] National Cancer Institute. Skin Cancer (Including Melanoma)—Patient Version; 2018. Accessed: 2018-03-22. Available from: `https://www.cancer.gov/types/skin`.

[8] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115.

[9] Waltz E. Computer Diagnoses Skin Cancers: Deep learning algorithm identifies skin cancers as accurately as dermatologists. IEEE Spectrum; 2017. Accessed: 2018-03-03. Available from: https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/computer-diagnoses-skin-cancers.

[10] Chan B. Solar lentigo; 2014. Accessed: 2018-04-22. Available from: https://www.dermnetnz.org/topics/solar-lentigo/.

[11] National Cancer Institute. What Is Cancer?; 2015. Accessed: 2018-03-03. Available from: https://www.cancer.gov/about-cancer/understanding/what-is-cancer\#tissue-changes-not-cancer.

[12] Pathan S, Prabhu KG, Siddalingaswamy P. Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review. Biomedical Signal Processing and Control. 2018;39:237–262.

[13] The Skin Cancer Foundation. Skin Cancer Information; 2018. Accessed: 2018-04-25. Available from: https://www.skincancer.org/skin-cancer-information.

[14] Kanimozhi T, Murthi A. Computer Aided Melanoma Skin Cancer Detection Using Artificial Neural Network Classifier. Journal of Selected Areas in Microelectronics (JSAM). 2016;8(2):35–42.

[15] Hameed N, Ruskin A, Hassan KA, Hossain M. A comprehensive survey on image-based computer aided diagnosis systems for skin cancer. In: Software, Knowledge, Information Management & Applications (SKIMA), 2016 10th International Conference on. IEEE; 2016. p. 205–214.

[16] Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain. 2008;8(6):221–223.

[17] Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: a tutorial. Computer. 1996 March;29(3):31–44.

[18] Kim P. MATLAB Deep Learning With Machine Learning, Neural Networks and Artificial Intelligence; 2017.

[19] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper with Convolutions. 2014 September;.

[20] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. 2017 May;60(6):84–90.

[21] Dai W, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Proceedings of the 24th international conference on machine learning. ICML '07. ACM; 2007. p. 193–200.

[22] Hoo-Chang Shin HR, Roth I, Mingchen Gao D, Le Lu RM, Ziyue Xu RM, Nogues RM, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. Medical Imaging, IEEE Transactions on. 2016 May;35(5):1285–1298.

[23] Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN Features off-the-shelf: an Astounding Baseline for Recognition. 2014 March;.

[24] Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning Deep Features for Scene Recognition using Places Database. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in Neural Information Processing Systems 27. Curran Associates, Inc.; 2014. p. 487–495.

[25] Wu N. Inception in TensorFlow; 2017. Accessed: 2018-03-20. Available from: `https://github.com/tensorflow/models/blob/master/research/inception/g3doc/inception_v3_architecture.png`.

[26] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2015 December;.

[27] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 2016 February;.

[28] Lin M, Chen Q, Yan S. Network In Network. 2013 December;.

[29] Nylund A. To be, or not to be Melanoma: Convolutional neural networks in skin lesion classification; 2016.

[30] Ridell P, Spett H, Herman P, Ekeberg Ö. Training Set Size for Skin Cancer Classification Using Google's Inception v3; 2017.

[31] ISIC-archive; 2018. Accessed: 2018-03-21. Available from: https://isic-archive.com/.

[32] DermQuest; 2018. Accessed: 2018-04-17. Available from: https://www.dermquest.com/.

[33] Dermatology Atlas; 2018. Accessed: 2018-04-17. Available from: http://www.atlasdermatologico.com.br/.

[34] Dermaamin; 2010. Accessed: 2018-04-17. Available from: http://www.dermaamin.com/site/.

[35] Dermoscopy Atlas; 2007. Accessed: 2018-04-17. Available from: http://www.dermoscopyatlas.com/diagindex.cfm.

[36] Global Skin Atlas; 2005. Accessed: 2018-04-17. Available from: http://www.globalskinatlas.com/.

[37] Hellenic Dermatological Atlas; 2011. Accessed: 2018-04-17. Available from: http://www.hellenicdermatlas.com/en/.

[38] Medscape; 2018. Accessed: 2018-04-17. Available from: https://reference.medscape.com/.

[39] Regional Derm; 2016. Accessed: 2018-04-17. Available from: http://www.regionalderm.com/.

[40] Skinsight; 2018. Accessed: 2018-04-17. Available from: https://www.skinsight.com/.

[41] PH2 database; 2012. Accessed: 2018-03-21. Available from: https://www.fc.up.pt/addi/ph2%20database.html.

[42] Thrun S. thrunlab/diagnostics; 2015. Accessed: 2018-04-27. Available from: https://github.com/thrunlab/diagnostics.

[43] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2016 March;.

[44] Bogley W, Robson R. Finding the Largest Inscribed Rectangle; 2018. Accessed: 2018-04-28. Available from: `https://oregonstate.edu/instruct/mth251/cq/Stage8/Lesson/rectangle.html`.

[45] Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine. 2013;4(2):627.