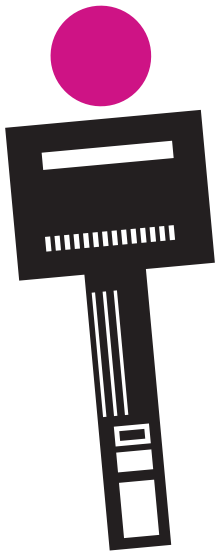


RAM ROBOTICS AND MECHATRONICS

Segmentation and Classification of Skin Lesions Using Neural Networks

J. (Jeffrey) Dokter

MSC ASSIGNMENT

Committee:
dr. ir. F. van der Heijden
dr. F.J. Siepel
prof. dr. ir. W. Steenbergen

August, 2020

035RaM2020
Robotics and Mechatronics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | TECHMED CENTRE UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Summary

When treating diseases involving skin lesions, properly assessing the severity of the disease is crucial in establishing a correct treatment plan. To do this, physicians classify cases based on guidelines. These guidelines often are a trade-off between consistent and quick assessment of the case. This leaves room for improvement on both sides.

In order to speed up the assesment process and make it more consistent, the possibilities of neural networks are explored. There are two ways in which neural networks are used to analyze images of skin lesion: segmentation and classification. Segmentation is used to detect and to localize the lesion area within the image. It is commonplace in medical image research and has been done on skin lesions before. Classification is used to indicate the severity of several aspects of a lesion or the disease itself. This is much rarer and mostly not in line with existing standards used by physicians, such as the ABCDE score for skin cancer and PASI score for psoriasis.

The goal of this research is to explore whether neural networks can be used to classify skin lesions in line with existing medical standards. Segmentation is also used to try and support the classification process. Due to the availability of data, the segmentation will focus on image of lesions regarding skin cancer, while classification focuses on psoriasis lesions and the corresponding PASI score. While these two parts are not directly connected, they explore adjacent applications. In the end, it is examined to see whether these two applications could be used together to improve classification results in the future.

While the classification results weren't good enough to indicate that this method is working and can be used in the future, there are enough changes that can be made to try and improve the results. Furthermore, the segmentation results are very promising, providing a way to aid the classification process by emphasizing the areas of interest from a image.

Contents

1 Introduction

Skin diseases are pretty common conditions which burden the suffering patients quite heavily. A survey showed they are the 18th leading cause of health burden worldwide, the 4th leading cause of non fatal burden, with three skin conditions listed in the 10 most prevalent diseases (Hay et al., 2014).

Te treatment of skin diseases greatly depends on the correct assessment of the severity of a case. Diagnosing skin diseases is mostly done with guidelines. However, a lot of these guidelines, such as the ACBDE method for skin cancer (Tsao et al., 2015) and PASI for psoriasis (Fredriksson and Pettersson, 1978), consider parameters that are not easily quantifiable, such as coloring or scaling. This results in diagnosis taking a lot of time in order to do it correctly, with PASI scoring taking up to half an hour. If the guidelines try to speed up the process, this leads to less accurate results. Also, parameters that are not easily quantifiable can lead to subjective assessments, since physicians can interpret those differently.

This research originally intended to try and solve this subjectivity issue. By using machine learning an algorithm could be trained to always provide the same diagnosis when similar cases are tested. This could also speed up the diagnosis process, since it is a simple operation of putting data into the algorithm instead of thoroughly examining the patient.

Since this would focus on skin diseases and skin cancer in particular, it would be easy to use image material as data for machine learning. This omits the need to extract features, which, as explained earlier, could lead to subjective data or might be hard to quantify.

Machine learning is used for skin lesion analysis in two ways: to detect a lesion within an image using segmentation and to determine the severity of the disease using classification. The purpose of this project would be to combine those two methods in some sort of pipeline. The image would first be segmented and the segmentation could then serve as an aid to the classification, by telling the classification where the lesion is and making the algorithm focus on that part of the image. This would be done for skin cancer, since most of the lesion segmentation work already done is focused on skin cancer (Maglogiannis and Doukas, 2009), making the segmentation step easier.

However, when it came to the classification, it turned out to be very difficult to obtain a large enough data set created by medical professionals. Fortunately, an opportunity arose with the Raboud UMC from Nijmegen. They were starting a project where machine learning is used to classify psoriasis. Unfortunately, they did not have any segmentation data available, so there wa no way to transfer the segmentation work already done to psoriasis cases. While this does create a disconnect between the segmentation part and classification part of this research, this seemed a good way to at least do research on skin lesion classification an try to see how segmentation and classification could work together, even though it might not be directly tested. This means the report is split in two seperate parts: the first one detailing research on how segmentation of skin cancer lesion and whether this could aid classification in the future and a second part describing an attempt to classify psoriasis lesion. In the end a conclusion will try and tie the two parts together to see whether the combination of these machine learning methods can be useful in the future.

2 Part 1: Segmentation

2.1 Introduction

Over the last couple of years, the occurrence of melanoma and skin cancer in general has increased. In The Netherlands alone, the amount of new incidents has increased from 3895 in 1990 to 16182 in 2017 (IKNL, 2018), which is more than four times as much. Studies show that when diagnosed early, the survival rate for skin cancer patients is way higher than when the cancer has spread through the body (98% to 18% (American Cancer Society, 2017)) . This makes it of the utmost importance that malignant skin lesions are found and treated as soon as possible. The 5-year survival rate also varies greatly, between above 90% in for instance The Netherlands, Denmark and Belgium and below 60 % in China, Ecuador and Taiwan(Allemani et al., 2018). That the survival rate and moment of detection are connected is confirmed even more strongly by (de Vries et al., 2004), where the thickness of newly discovered melanoma is higher in eastern European countries than in western European countries. This corresponds with the numbers in (Allemani et al., 2018), where western European countries all have a survival rate above 80%, while almost all eastern European countries are below 80%, with the exception of The Czech Republic. In short, skin cancer is a disease with a growing amount of incidents where survival greatly depends on early diagnosis and treatment.

For current diagnosis, a dermatologist is involved to examine whether suspicious skin lesions are malignant or not. When a skin lesion is determined to be malignant, the dermatologist classifies the severity of the skin cancer with just the naked eye, based on the ABCDE rule: Asymmetry, Border, Color, Diameter and Evolution (Tsao et al., 2015). This means that while there are some parameters that can be checked, there is always some kind of subjective part during the diagnosing. This is very unsettling, because the treatment will greatly depend on the severity of the skin cancer, just as the expected survival of the patient themselves, as explained earlier. To increase the rate at which patients are examined and in order to improve the objectivity of the classification result, a method not using human interaction would be preferable. This means that instead of having a physician or another qualified person look at the lesion, a non-human entity inspects it. For skin lesions, this means analyzing image data, since images are the most common and easiest way to convert information from skin lesions into data.

Analysis on image data has already been a big part of scientific research in the past, and therefore analyzing data on skin lesions has also been done before, some of the first literature dating back to 1987 (Cascinelli et al., 1987). For the research by Cascinelli et al. (1987), color slides were digitized and analyzed. This approximates the use of an image taken by a digital camera as is common now. However, there are more non-invasive imaging techniques for skin lesions: dermatoscopy, multispectral imaging, laser-based enhanced diagnosis, optical coherence tomography, ultrasound imaging and magnetic resonance imaging. Off all of these techniques, only dermatoscopy is also used regularly (Masood and Al-Jumaily, 2013). This means that there are two commonly used types of images available for analysis:

- dermatoscopic images are made using a dermatoscope, an instrument specifically designed to get skin related information. Traditionally, it uses a liquid medium to take away skin surface reflections, but more recent versions use polarized light instead.
- non-dermatoscopic images are made using regular digital cameras and are just a more limited substitute for looking at a lesion with the naked eye, although some unwanted artifacts might be introduced.

It has been proven that using dermatoscopy during skin lesion diagnosis improves accuracy (Bafounta et al., 2001)(Carli et al., 2004)(Kittler et al., 2002)(Vestergaard et al., 2008)(Westerhoff et al., 2000). However, these studies also show that in order for the accuracy to improve, the user has to be trained, because non-trained users perform the same with dermatoscopes as with the naked eye (Westerhoff et al., 2000). Furthermore, in order for dermatoscopy to be performed, a dermatoscope has to be purchased as well. While this may not be an expensive purchase, it still is an additional investment. Requiring both a trained operator and an additional piece of equipment also forces patients to again go to the hospital in order for images to be taken. This means human interaction is required again and limits possibilities for patients to do any preparation or data gathering themselves.

Non-dermatoscopic imaging only requires a digital camera, which most people own, either within a smart phone or separately. So using digital images for analysis makes the system more accessible, especially with the future, where patients uploading their own data is a possibility, in mind. Therefore, this research will try to focus on the analysis of non-dermatoscopic images.

Because this problem has been researched during a long time period already, the techniques for the analysis has also varied greatly. The very first paper by Cascinelli et al. (1987) used histograms based on different color spectra to determine threshold values, after which they could apply a contrast on the image separating the lesion from the other skin. However, this very basic method can only be applied after researching the particular image and then selecting the thresholds, making this method slow and not generalizable. The research by Cristofolini et al. (1997) already uses a fully automated system, using techniques like edge enhancement, shape evaluation and color analysis to diagnose skin lesions. While the way the system works is more promising with regard to the desired implementation, it does not achieve the same results a trained dermatologist would have.

Starting at the end of the 2000s, computer analysis is way more common and advanced, as can be seen in the overviews by Maglogiannis and Doukas (2009) and Masood and Al-Jumaily (2013). Furthermore, the analysis process is now divided into several components: preprocessing, segmentation, feature extraction, feature selection and classification. This also means that not everything has to be done using computerized techniques. A physician could, for instance, segment an image by hand, after which an algorithm extracts and selects features and classifies the image. Furthermore, machine learning is more commonly used now, in the form of both neural networks (NNs) and support vector machines (SVMs).

Machine learning is a form of algorithm development where the algorithm is progressively trained instead of explicitly programmed. This means the focus is more on the output of the algorithm instead of the inner workings, since those are not determined and programmed by hand, but by the training. Both NNs and SVMs are methods for supervised learning. This means that there is input data with desired output data available to train the algorithm. So the training is done by comparing the actual output with the desired output and making adjustments to the parameters such that the difference, or loss, is minimized. For the segmenting a lesion images, for instance, the input would be a full colour picture containing the lesion, while the desired output would be black and white picture highlighting the lesion.

Support vector machines are a methods used in classification analysis. It can separate a dataset into two categories given examples of both categories. If the data is visualized as a space, a support vector will draw a line or plane between the two categories. It will do this with as large a gap as possible. This means it can find the global minimum for the loss based on

the training data that has been provided, which is a big advantage of using SVMs. However, since SVMs can only distinguish two categories, it must be chained multiple times if data has to be classified in more than two. This is not a problem with regard to segmentation, since a pixel either belongs to the lesion or does not, but it can be hard when classifying lesion into categories 1 to 5 based on severity. But, the even bigger problem is that SVMs can only classify one data point simultaneously if only one SVM is running. This means that an SVM cannot looks at groups of pixels in a photograph, losing important information inherent to images like textures and features such a edges or shapes. These can only be examined when the area the algorithm looks at, or receptive field, is a group of pixels. This makes SVMs less suited for image processing, especially when textures are important.

Neural networks are algorithms based on human brains. They consist of neurons which are activated based on a sum of all inputs. Those inputs are weighted, the result is biased and sent to the output. The neurons are structured in layers, with every layer that is not the input or output layer being called a hidden layer. The amount of neurons in every layer can be designed the way the user wants to. This means a neural network can accept all kinds of input, produce all kinds of outputs and be configured in the hidden layers just the way the user wants, making it very flexible. Furthermore, they consist of techniques that make the composition of the connection between layers different as well. A standard neural network would have connections between all neurons in subsequent layers, but there are other types of layers requiring less connections, like convolutional or pooling layers.

Convolutional neural networks(CNNs) use both convolutional and pooling layers to approach the data processing from a more image processing perspective. Both convolutional and pooling layers are summarizing information from multiple input signals into a single output signal, effectively grouping information. This means that when an image is the input of the network, information about separate pixels is grouped, increasing the receptive field of a subsequent layer in the network. This means convolutional networks can look at textures and shapes, giving them an edge over SVMs and regular NNs in image processing. This is why these networks are used for this project.

Using CNNs for lesion segmentation certainly has been done before (Nasr-Esfahani et al., 2017)(Badrinarayanan et al., 2017)(Shelhamer et al., 2017), and this research is not trying to figure out a way to improve that with a whole new architecture. This research is done to answer the question: can segmentation done by convolutional neural networks be used to aid skin lesion classification?

2.2 Methods

2.2.1 Neural Network Architectures

As mentioned in Section 2.1, research with regard to image segmentation has already been done. This means that for this project, existing methods will be used. So no new design or design elements will be introduced for the neural network doing the segmentation. However, even though this problem has been mostly solved, this has been done in many different ways. Therefore, in order to achieve optimal results, different solutions are examined, after which one design is chosen. The designs examined here are U-Net (Ronneberger et al., 2015) and dense fully convolutional networks (DFCN) (Nasr-Esfahani et al., 2017). Other designs have been researched, but were deemed unfit for this assignment because they prefer smaller resource use and faster performance over detailed results or are less recent.. Because the extraction of the features depends highly on the accuracy of segmentation, less detailed results are detrimental to the other parts of this research. Other options that were considered were: SegNet (Badrinarayanan et al., 2017), fully convolutional networks tested with Jaccard distance

loss(Yuan et al., 2017), FCN-8s (Shelhamer et al., 2017) and dilated convolutions (Yu and Koltun, 2015).The unique traits of each network are discussed in the following subsections.

U-Net

In a regular convolutional network, using pooling and convolutional layers will increase the field of view of neurons in deeper layers. The field of view is the aspect making convolutional network differ from regular neural networks, because it looks at structures within an image, by grouping values. However, this comes at the cost that a lot of details are lost, since information about groups of neurons is packed into one single value, like an aggregation. This makes the effective resolution of the data out of every layer lower. So even if there is a segmentation map at the output, it will be coarse. However, the extended field of view is necessary to detect larger features within an image.

U-Net (Ronneberger et al., 2015) tries to solve this loss of details by adding feature channels to the network. During the convolutional path, everytime the data is about to be pooled, a snapshot of the data is taken and saved. After the convolutional path, a deconvolutional path is executed, where the data is upsampled and interwoven with the snapshot of the according resolution. This result in the details being added back into the data, while keeping the data from the largest field of view. An image of the network structure is shown in Figure 2.1 Although the

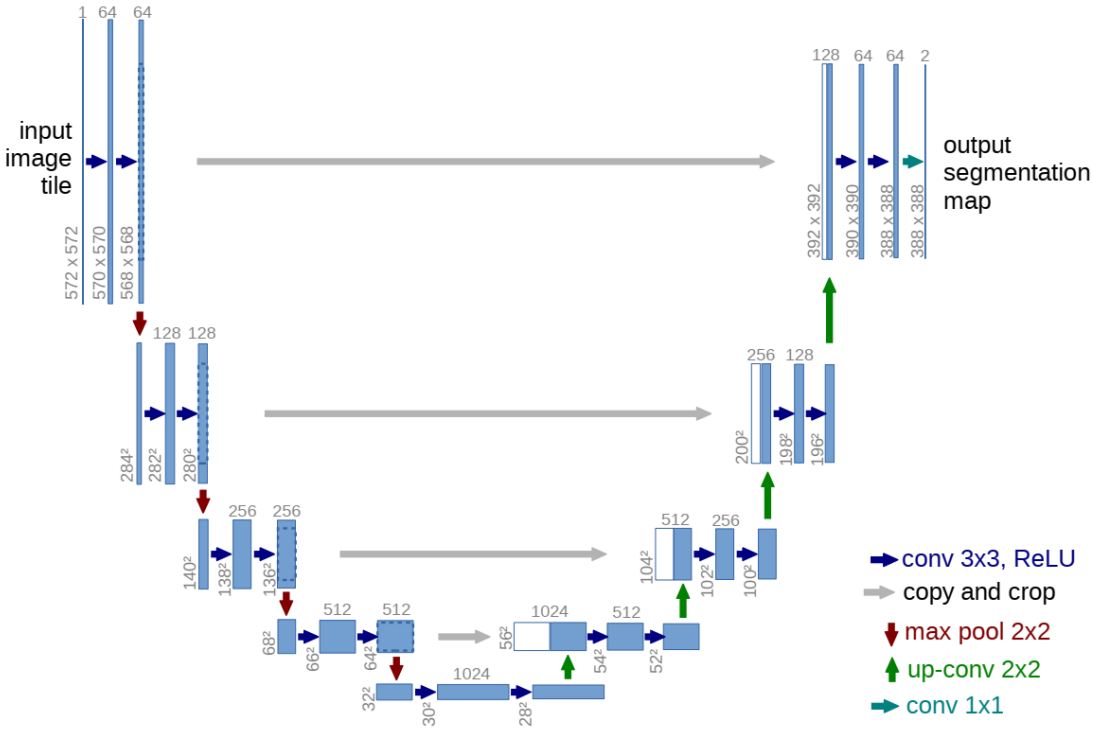


Figure 2.1: U-Net network structure (Directly taken from Ronneberger et al. (2015))

paper describing U-Net is from 2015, which is old in respect to the alternatives, it is still used and improved upon. It also has been used widely, so several implementations in different code bases and using different frameworks exist, so examples can be used as a starting point. It also strictly consist of standard layers, making it very easy to implement in any coding language and using any framework. Even though this network was designed to segment biological tissue in a binary manner, it was not designed for skin segmentation specifically. This means that if necessary

adjustments to the network design have to be made. Furthermore, it is probably better to train the network from scratch instead of using already existing weights as a starting point.

DFCN

Dense fully convolutional networks (Nasr-Esfahani et al., 2017) try to solve the loss of information in pooling layers very differently from U-Net. It introduces dense pooling layers, which do not throw the unused information away, but instead feeds it into a mirror network equal to the network the chosen values are fed into. At the end of those networks, the results of the mirror networks are interleaved. Because of this interleaving process, no deconvolution in necessary.

DFCN were designed specifically for skin segmentation, also performing better than U-Net, as shown in the original paper. However, the interleaving is done in special Dense Pooling layers designed by the researchers themselves. An example network layout including Dense Pooling layers for 1D is shown in Figure 2.2 Even though the code and math for these layers is available, adapting those layers to another platform or code base is going to be more work than using standard layers.

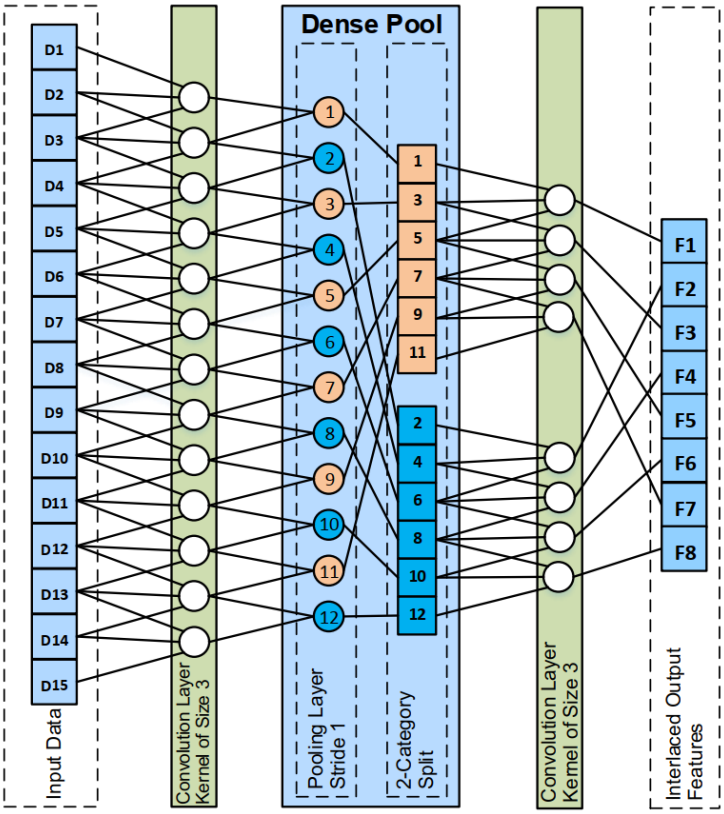


Figure 2.2: DFCN example 1D network layout. (Directly taken from Nasr-Esfahani et al. (2017)

Decision

Due to a lack of experience in working with neural networks, it is preferable to work with a network that is easy to implement in any coding language with any framework. That way a code and framework can be freely chosen based on preference and ease of use. Furthermore, the difference in performance shown in the paper by Nasr-Esfahani et al. (2017) is within one percent for most metrics, making the trade off between performance and ease of use not worth choosing the better performing network. Therefore the decision was made to start with implementing U-Net to get familiar with implementing and training a neural network and at least have an

option for segmenting pictures. When time allows it, DFCN can also be implemented to allow for comparison between the two different networks.

2.2.2 Data

For skin lesion segmentation of skin cancer lesions, there is already data available to train, validate and test the networks. The data used for this assignment is from the ISIC archive (The International Skin Imaging Collaboration: Melanoma Project, 2019), an archive of images and segmentation collected by The International Skin Imaging Collaboration: Melanoma Project it contains close to 24000 images and with almost all accompanied by segmentations. Those segmentation or on two levels: expert and novice. These images comes from several different datasets. While the focus for this assignment is on regular digital imaging besides dermoscopy, the images from this dataset are all dermoscopic. This is unfortunate, but acquiring data using regular photography would require patients, as well as physicians to do the segmentation, which is something that would take too much time and is out of scope for this assignment. Furthermore, while this data does not cover the entire spectrum of images that is desired, it provides a good starting point to train the networks. A subset of 5000 images is used for training and validation, with this set split 70/30 between training and validation. So 3500 images are used for training and 1500 for validation. Finally, another 100 images were used to test the network when training was done.

In order to somewhat prevent these problems, a small dataset is used in conjuction with the ISIC data. This data set consists of 63 images, which are all just regular digital images. Therefore these images contain common effects like light flares and uneven lighting in general. This dataset is from the Vision and Image Processing Lab from The Univeristy of Waterloo (2019). This dataset is used as evaluation data during the training process, to see whether the network can also segment regular digital images properly.

Implementation

Even though the network to be designed has been decided, there are still several factors to be chosen. First off a coding language has to be chosen to use. For this Python is used, since it is very easy to use and has a wide array of both examples and libraries that can be used to skip a lot of the detailed work, making the implementation faster. Then the library to work with was determined to be Tensorflow. It is aimed toward machine learning and optimizes computations before running them. This and the fact that the library is optimized for running on different hardware platforms, like GPU, makes it very efficient to work with. Furthermore, it has the Keras library that functions on top of it, providing even higher level access. Keras is an alternative if working with Tensorflow directly proves to be too challenging. In the end, this turned out to be the case, so the network was implemented with the Keras library.

Since the images used to train and evaluate the network are of a different resolution than the 572x572 used in the original U-Net paper (Ronneberger et al., 2015), the layers were reshaped to fit the images better. The input images all have a size of 592x400, so the input layer will also have this size. The network is trained using both SGD and Adam optimizers.

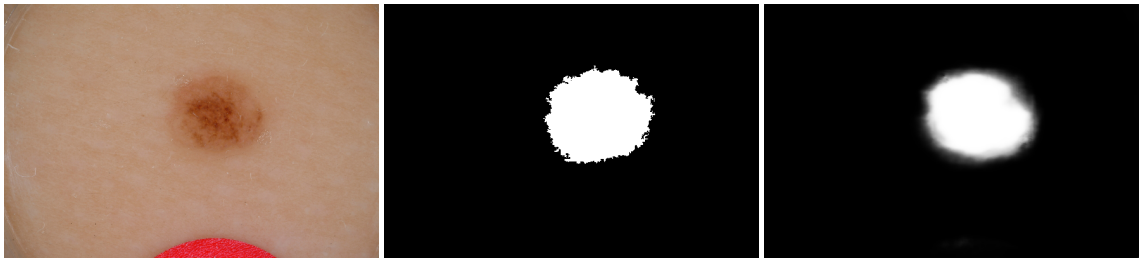
Because training this network is not feasible on a regular computer, an alternative has to be found. Fortunately, the people from the DSI (formerly CTIT) made their cluster available to be used for this project. This way, the network can be trained more effeciently. Since the cluster contains Nvidia GPUs, the CUDA version of Tensorflow is used underneath the Keras library in order to lower runtimes of the code.

2.3 Results

During training the best result obtained was 98.36 % accuracy, with a loss of 0.0383 on the training data. On the evaluation data, the accuracy was 98.23 % and the loss 0.2787. The weights obtained from this training run were used during the continuation of the process. The results of the testing set were an accuracy of 98.33 % and a loss of 0.0746. Some additional metrics were calculated for the testing dataset, such as the sensitivity (true positive rate), specificity (true negative rate), precision (positive predictive value), miss rate (false negative rate) and fall out (false positive rate). How these values are calculated can be seen in Appendix A. These metrics were calculated for every image and then averaged for all 1000 images in the dataset. The results were:

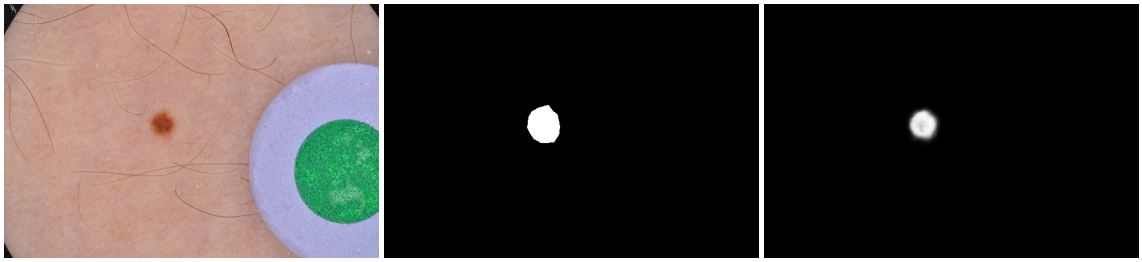
- Sensitivity: 0.69
- Specificity: 0.99
- Precision: 0.83
- Miss Rate: 0.31
- Fall-out: 0.01

Some examples of segmentations from the network are shown in Figure 2.3, Figure 2.4.



(a) Image **(b)** Intended Result **(c)** Actual Result

Figure 2.3: Example of skin segmentation results



(a) Image **(b)** Intended Result **(c)** Actual Result

Figure 2.4: Example of skin segmentation results

2.4 Discussion

First of all the consistency of the accuracy across the training, validation and test data sets of approximately 98.3 % shows that the network is not overtrained for the training data, which is very promising.

While the network is almost always able to find the correct location of the lesion, the details are not very clearly defined in the results. Especially when the lesion fades at the edge, they are incorrectly segmented. The effect of this can be clearly seen in the difference of accuracy between Figure 2.3 and Figure 2.4, where the border of the second lesion is better defined than that of the first. However, the results in Figure 2.4 also clearly show that the network does not

pick up on disturbances, such as the hairs or the pad that can bee seen in the photo.

The additional metrics support this. So while the sensitivity is low and the miss rate is high, the specificity is high and the fall out low. A low sensitivity and high miss rate indicate that there are relatively little true positive and relatively many false negatives, meaning that a lot of the pixels that should be in the lesion, are not recognized as such. However, the high specificity and low fall out mean there are relatively many true negatives and relatively few false positives. So the network does not pick up all the pixels belonging to the lesion, but when it indicates a pixels as part of the lesion, it is nearly always correct.

While these results prevent a diagnosis that is more severe than the actual case, the opposite is somewhat likely to occur. This means the assessment could be that the lesion is less severe than it actually is, which is just as problematic. A way to solve this is to train to network to be more aggressive. This means increasing the learning rate and trying to find and even lower minimum in the loss curve than it currently holds.

However, if this segmentation information is used in the following classification process, there are several ways to work around the possible inaccuracies discussed above. By assuming that the segmentation is not perfect, the input image for the classification network can be weighted according to the segmentation, instead of using a mask. Using a mask could reduce pixel values of the input image to zero, resulting in a chance that a pixel that is part of the lesion is not used for classification. This can be prevented by offsetting the segmentation slightly, meaning a 0 will be a small value, but not zero. This means that pixels that are not segmented as part of the lesion are diminished but not taken out of the classification, making every pixels part of the following step. Furthermore, the high specificity means that few of the pixels that are not part of the lesion, are falsely amplified during the classification. Also, in order to calculate the additional metrics every pixel needs to be true or false. This is accomplished by thresh holding every pixel at 0.5. This means that the segmentation is more nuanced than the additional metrics. This means that weighting might not even be necessary, since there will be very few pixels that are 0 within the segmentation, only very close to 0.

2.5 Conclusion

While it is hard to say that U-Net is definitively the best architecture to use when segmenting skin lesion of skin cancer, it provides acceptable results, with an 98 % accuracy over the 1000 test images. This means that this network could be used to alter input images for a classification network to aid that process.

2.5.1 Recommendations

In order to improve the segmentation accuracy, some additional training could be done, focusing on resolving the biggest flaw demonstrated in the results, the lack of sensitivity.

Another option is to try another network architecture. While U-Net is widely used and has proven to be effective in a variety of situations, there are other options. DFCN (Nasr-Esfahani et al., 2017) shows great promise, but is more difficult to implement. Other options such a SegNet (Badrinarayanan et al., 2017) are also available.

When actually using segmentation as an aid for classification, care should be taken that the segmentation fits the input layer of the classification network, such that no information is lost or artificially added. Finally, keep in mind that segmentation is just an aid and as long as it is not 100 % accurate, it should not fully determine what the classification takes into account.

3 Part 2: Classification

3.1 Introduction

Psoriasis is a skin disease the manifests in the form of lesions on the skin. These lesion can be red and scale to various degrees. It is quite common, with some source reporting 4.5 million cases in the United States alone (Stern et al., 2004). A lot of these patients also state psoriasis to be a large problem in every day life (Stern et al., 2004). On top of that, psoriasis cost were estimated to be close to be between 1.6 and 3.2 billion USD in 1993 (Sander et al., 1993). This number has probably only increased, since reports show an increase in the amount of cases globally (Organization, 2016). The amount of discomfort is also proportional to the severity of the disease and patient are not always satisfied with treatment plans (Stern et al., 2004). Therefore, correct diagnosis of the severity of the disease is paramount.

Currently, the most common way to asses a psoriasis case is using the PASI scoring system (Fredriksson and Pettersson, 1978). However, using this system has a lot of issues and diagnosing a single patient takes approximately half an hour. Furthermore, it takes severity parameters, such as redness as does not assign them quantifiable scores, making the scoring process vulnerable to subjectivity.

In an effort to make the classification of psoriasis severity more objective and faster, the Radboud UMC started a project to try and use machine learning for classification. This research will help as some investigative work for a larger project to try and figure out the problems that will be faced when trying such a type of solution.

There are many machine learning methods, but for this project the method of use will be convolutional neural networks(CNNs). They have a big advantage when images need to be analyzed and there is already familiarity with this method, making implementing the experiments easier. For an extensive comparison between CNNs and other machine learning methods, please refer to Section 2.1.

Some work has already been done with regard to PASI classification and machine learning, with research done by Pal et al. (2016) showing promising results. However this research does not use psoriasis images directly as input for a CNN, but extracts features from the images and uses does. More direct approaches such as done by Zhao et al. (2020) show worse results. This means that there is much room for improvement.

This research is a precursor for a bigger project started between the University of Twente and the Radboud UMC and is more explorative. The main goal is to figure out the main problems that will present itself when trying to classify psoriasis using neural networks and machine learning in general. It will try to answer the question: Is it feasible to use convolutional neural networks for PASI classification of psoriasis lesions?

3.2 Background

3.2.1 PASI

The Psoriasis Area and Severity Index (PASI) was introduced in 1978 as a new way to asses psoriasis cases (Fredriksson and Pettersson, 1978). It has become the standard since then. It uses the Area and three severity factors to asses the grade of the disease:

- Desquamation (scaling)

- Induration (thickness)
- Erythema (redness)

Area

The Area property of the PASI is the most straightforward one: more area covered in lesions, means a higher score. The score goes from 0 to 6 with the scores representing a percentage of area covered:

- 0: 0 %
- 1: <10 %
- 2: 10-29 %
- 3: 30-49 %
- 4: 50-69 %
- 5: 70-89 %
- 6: 90-100 %

Severity Parameters

The three severity parameters are all scored using the same method: from 0 to 4 with the indicators of the scores being absent, mild, moderate, severe and very severe. Representations of each parameters and corresponding score can be found in Figure 3.1, Figure 3.2 and Figure 3.3.

(a) Absent **(b)** Mild **(c)** Moderate **(d)** Severe **(e)** Very Severe

Figure 3.1: Example of desquamation scores.(DermNet NZ, 2020)

(a) Absent **(b)** Mild **(c)** Moderate **(d)** Severe **(e)** Very Severe

Figure 3.2: Example of induration scores.(DermNet NZ, 2020)

The severity parameters are the biggest problem with regard to objectivity. While the area score

(a) Absent **(b)** Mild **(c)** Moderate **(d)** Severe **(e)** Very Severe

Figure 3.3: Example of erythema scores.(DermNet NZ, 2020)

represents an percent range, the severity score only is some kind of indication. However, the difference between for instance mild and moderate can be hard to distinguish in some cases. There is no clear cutoff between two severity scores. This leaves scoring these parameters open

for interpretation of the physician, which in turn can result in different scores for the same case based on who is scoring it.

Score Calculation

To calculate the final score, the body is separated into four areas: the head, the arms, the legs and the body. For each area each parameter is determined. The three severity parameters are summed per area and multiplied by the area score. Then the scores are weighted based on the area: 0.4 for the legs, 0.3 for the body, 0.2 for the arms and 0.1 for the head. Finally, the four scores are summed, resulting in a score between 0 and 72.

3.3 Methods

3.3.1 Neural Network Architecture

Image classification has been a staple machine learning problem for a while, especially when using convolutional networks. Since convolutional networks are suited for texture recognition and larger feature detection, a plethora of classification problems have been mostly solved using them. This also means the a large selection of network architectures is available in literature to try and use for this research.

From all of the available architectures, VGG16 (Zhang et al., 2016) is one of the most well known and widely used. The network consists of blocks of three layers. Each block has two convolutional layers followed by a pooling layer. The convolutional layers have a filter size of 3x3 and the pooling layers do the same with a stride of two, resulting in each block decreasing the size in half in both dimensions. The final layers are fully connected with at the end a softmax layer. The whole architecture is shown in Figure 3.4.

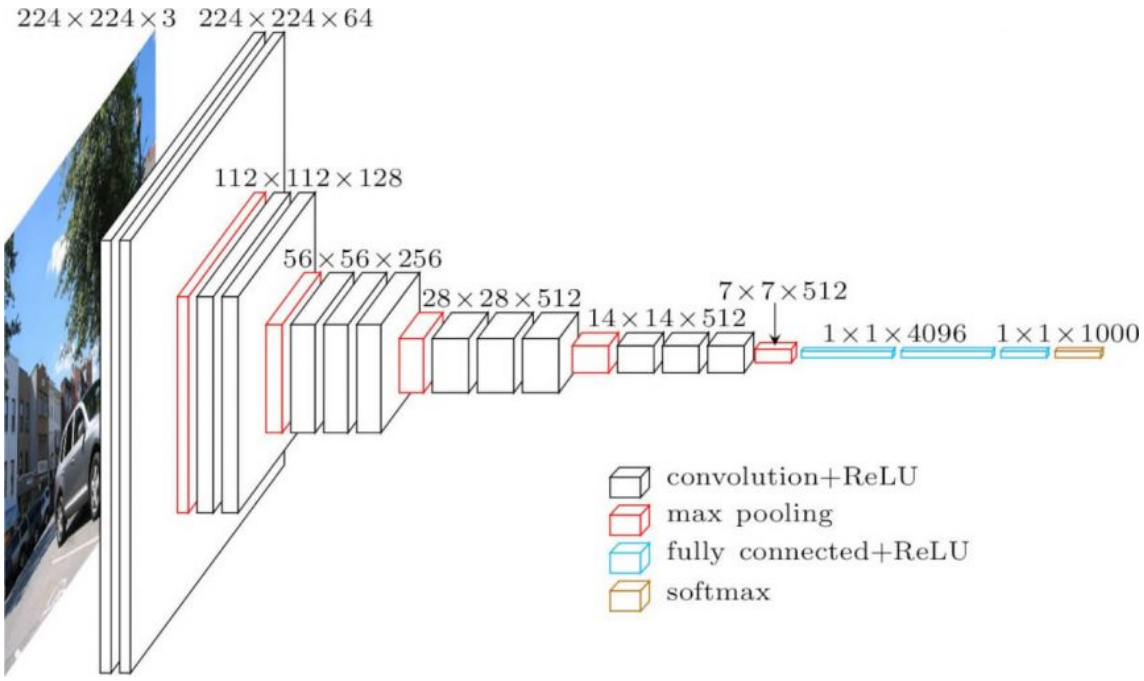


Figure 3.4: VGG16 Architecture

Since the VGG16 architecture was based on research done by **?**, and it has proven to work in a variety of classification problems, the decision was made to base the network architecture used for this research on VGG16.

3.3.2 Data

There is next to no clincal image data of psoriasis lesions publicly available like there are for skin cancer. Furthermore, this data lacks corresponding PASI classifications. Fortunately, there was an oppurtunity to work with the Radboud UMC on this project. The Radboud UMC has a database consisting of clinical color images of psoriasis along with PASI scores. In order for the images to be widely used, they needed to be properly labeled. Also, to ensure the privacy of the patients, some images needed to be cropped to remove facial and other recognisable features such as prominent moles, scars and tattoo's. This was all done during this project, also to support future projects with a complete and ready-to-use data set.

During this project, another group of researchers from the Radboud UMC started with a project that more widely looks at machine learning with respect to psoriasis and PASI scores. After discussing with this group, the decision was made that they will focus on images of the body, this research will focus on images of the arms.

When implementing a neural network to classify a psoriasis photo to a PASI score, a big problem becomes immediately clear: PASI scores are determined for body parts as a whole. However, the corresponding body part cannot be entirely depicted in a single image. The arms, for instance need to have at least photos taken from both the front and the back and then it can be argued that the sides are not clearly visible when combining them as the image of the arms. In the case of the dataset used for these tests, there are four images of the arms: front and back of the left arm and front and back of the right arm. This means that for the combination of these four images, a singular score is available. The images cannot be separated, since in most cases not every photo contains the lesion material with properties that match the score. So the score is an average of what is visible on all of the images. This means that the use of single images with the score for all four images will not result in a properly trained network. Because it was not feasible to use advanced image processing techniques to circumvent this problem, the decision was made to concatenate all the images corresponding to one score together in a set way. The set order is introduced to get some consistency within the input data for the network, in an effort to make the training process easier. The order of the images from left to right, top to bottom is as follows: back left, front left, back right, front right.

While there are four different score in PASI, not all are easily classifiable using a single neural network. Since induration is more of a property in the third dimension, this is next to undistinguishable in an image. Area is also hard, because it is almost necessary to combine information on all the lesions in the image, instead of averaging. An additional problem is that while all the lesion areas need to be congregated, there can be overlap between areas in different images. This leaves erythema and desquamation. The project will focus on erythema, since this is coloring, which should be the most easy to recognize, provided color images are used.

In order for there to be consistency in the dataset used for the network, all four images of the arms need to be available for the corresponding erythema score to be put into the dataset. This resulted in 612 concatenated images and scores. In order to increase the dataset all of the images were mirrored and then concatenated, resulting in 1133 images total. This is due to some PASI classifications not having a score for the erythema, so some images did not have matching labels. This is divided with approximately 15 % for the test set and 15 % for the validation set and the remaining 70 % for the training set. So the final split for training, validation and testing images is 793, 170 and 170 images.

3.3.3 Implementation

The concatenated images described above have a size of 9600x6400. This is not compatible with the regular VGG16 layout, in which the input layer has a size of only 224x224. Reducing the image to the correct size for the network would result in a loss of detail in the image, while increasing the size of the input layer would either disturb the architecture or increase the amount of parameters too much. So a compromise was made to add a block of layers in front of the network of two convolutional layers and a pooling layer. These layers adhere to the same configuration as the other block in the network, but increase the size of the input layer to 448x448. The images are reduced to this size. The network is trained using SGD and Adam optimizers.

The network is coded in Python using Tensorflow with Keras on top of it. The reasoning behind this decision can be found in Section 2.2.2. In order to maintain the data security and patient privacy, the training of the networks was done on servers of the Radboud UMC itself. This meant the data always remained on secure servers of the medical facility and was accessed remotely. This was done with the help of the Diagnostic Image Analysis Group (DIAG).

3.4 Results

The results achieved with the best trained network were as follows:

	Loss	Accuracy
Training Set	1.67	0.493
Validation Set	1.79	0.224
Test Set	1.24	0.5

Table 3.1: Results of network training and evaluation

3.5 Discussion

What immediately stands out is that the accuracy is not consistent across all data sets. This is probably due to the size of the training set being relatively small. This mean that the network is not training to recognize a lot if diverse images, meaning that it can overtrain on the traning images. If the images of the test set closely resemble al lot of those of the training set, this means that the accuracy is close to that of the training set, while the of the validation set is significantly lower.

However another reason can be that the accuracy of the training is just not high enough. While approximately 50 % is better than randomly guessing, which would mean 20 % since there are five different classes, it is nowhere near high enough to reliably classify images. Therefore using the validation and testing sets has little values, since the learning on to training set is just not completed yet.

In the end, it seems that the training did not catch on enough, with small improvements in accuracy in loss, but nothing above the 50 % shown here. Since the training of neural networks is an autonomous process, the exact reason cannot be determined, but some guesses can de made.

The training process just did not catch on. This can be due to not using the correct parameters or just some though luck. This means that experimenting more with the learning rate, using different optimizers and just doing more and longer runs might deliver better results in the end.

The data is not suitable for being used in the network the way it was. The way the images are being fed into the network for these experiments results in a loss in detail. Furthermore,

the relevant parts of the image only make up a fraction of the image, The images do not only contain skin and lesion, but most also contain a significant amount of background. This means that with the image being reduced in size makes it too hard for the network to pick up on the more subtle differences between the scores. Using image reconstruction techniques to make a 3d reconstruction of the arms and then rolling the skin out to a 2d surface might be a way to circumvent the addition of clutter to the network, making sure it only has to focus on the skin and lesion tissue.

The network architecture was not matching with the purpose of the network. While the VGG16 architecture has been proven to work on a wide variety of classification problems, this does not mean it works on all of those problems. This means that the network is too shallow or too deep to get a proper result for this classification problem. Layer sizes can also be of influence. This ties back into the previous possible problem of the image details being lost when put into the network. Increasing the size of the input layer will diminish that problem, but can result in other problems, such as the amount of parameters being to large and messing up the VGG16 architecture.

3.6 Conclusion

With the available results it looks like it is not feasible to classify images of psoriasis lesions using neural networks. However, there are several changes that could be tried to see if there is another way to make this work. The fact that training neural networks is a process that does not provide complete insight into its inner workings makes it hard to figure out what the issues are. This also means that it is hard to rule it out as a possible solution to the classification problem. However, following these experiments, there are no clear signs that this might work without significant changes; either to the data, training process, network architecture or even machine learning method.

3.6.1 Recommendations

The clear issue is that the multiple images correspond to a single label. All of these images need to be taken into account for input to adhere to the score. This is probably a big issue, since redundant information will be added into the network, which will be at the cost of detail. Furthermore, because the entire are needs to be on the image, it is unavoidable that background is incorporated into the images as well, since no body area is rectangular. This introduces unnecessary information into the images, reducing the relevant area with respect to the images even more. So a possible solution would be to alter the data using image processing techniques to find a way to just show the skin surface as a single image. This might involve 3d reconstruction followed by rolling out the skin into a 2d surface again.

This training and testing experiment was quite limited in scope. So trying different network architectures and different training parameters could also help improves results.

It can also simply be that neural networks are not the machine learning tool suited to solve this issue. Alternatives, such as SVM can be tested as well.

4 Conclusion

While in the end, the segmentation training process proved to be quite successful and promising, it could not directly be applied to the classification process. However, looking at the issues found when trying the classification, it seems like segmentation could provide a means to improve results. The way the data for the classification network is being set up right now, the areas of interest for classification are sparse within the data. This means that providing the classification network with a means to focus on specific areas could by using segmentation could be ideal. Therefore, using the conclusions found in both parts, it is fair to say that segmentation and classification may certainly be used together to improve classification results and training processes in the future.

4.1 Recommendations

In order to further investigate this conclusion and make it more definitive there just needs to be some research done of both methods together. This can easily be achieved by either obtained classification data for skin cancer or segmentation data for psoriasis. It will be easiest to use segmentation data of psoriasis lesion to try and training the existing segmentation network using transfer learning. This is mainly because there are no real classification results anyway and this means there is some progress for segmentation training, while a classification network for skin cancer needs to be trained from scratch.

A Equations for metrics

A.1 Binary Classification (Segmentation)

Where T stands for true, F stands for false, P stands for positive and N stands for negative.

Sensitivity or true postive rate (TPR):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{A.1}$$

Specificity or true negative rate (TNR):

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{A.2}$$

Precision or positive predictive value (PPV):

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{A.3}$$

Miss rate or false negative rate (FNR):

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR} \tag{A.4}$$

Fall-out or false positive rate (FPR):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR} \tag{A.5}$$

Bibliography

Allemani, C., T. Matsuda, V. Di Carlo, R. Harewood, M. Matz, M. Nikšić, A. Bonaventure, M. Valkov, C. J. Johnson, J. Estève, O. J. Ogunbiyi, G. Azevedo e Silva, W.-Q. Chen, S. Eser, G. Engholm, C. A. Stiller, A. Monnereau, R. R. Woods, O. Visser, G. H. Lim, J. Aitken, H. K. Weir, M. P. Coleman and C. W. Group (2018), Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries, **vol. 391**, no.10125, pp. 1023–1075, ISSN 0140-6736, doi:10.1016/S0140-6736(17)33326-3.
http://www.sciencedirect.com/science/article/pii/S0140673617333263

American Cancer Society (2017), Cancer Facts & Figures 2017.
https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html

Badrinarayanan, V., A. Kendall and R. Cipolla (2017), SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, **vol. 39**, no.12, pp. 2481–2495, ISSN 0162-8828, doi:10.1109/TPAMI.2016.2644615.

Bafounta, M.-L., A. Beauchet, P. Aegerter and P. Saiag (2001), Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests, **vol. 137**, no.10, pp. 1343–1350.

Carli, P., V. D. Giorgi, E. Crocetti, F. Mannone, D. Massi, A. Chiarugi and B. Giannotti (2004), Improvement of malignant/benign ratio in excised melanocytic lesions in the 'dermoscopy era': a retrospective study 1997-2001, **vol. 150**, no.4, pp. 687–692, ISSN 1365-2133, doi:10.1111/j.0007-0963.2004.05860.x.
https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0007-0963.2004.05860.x

Cascinelli, N., M. Ferrario, T. Tonelli and E. Leo (1987), A possible new tool for clinical diagnosis of melanoma: The computer, **vol. 16**, no.2, pp. 361–367, ISSN 0190-9622, doi:10.1016/S0190-9622(87)70050-4.
http://www.sciencedirect.com/science/article/pii/S0190962287700504

Cristofolini, M., P. Bauer, S. Boi, P. Cristofolini, R. Micciolo and M. C. Sicher (1997), Diagnosis of cutaneous melanoma: accuracy of a computerized image analysis system (Skin View), **vol. 3**, no.1, pp. 23–27, ISSN 1600-0846, doi:10.1111/j.1600-0846.1997.tb00155.x.
http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0846.1997.tb00155.x

DermNet NZ (2020), PASI Score.
https://dermnetnz.org/topics/pasi-score/

Fredriksson, T. and U. Pettersson (1978), **vol. 157**, no.4, pp. 238–244, ISSN 0011-9075, doi:10.1159/000250839.

Hay, R. J., N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, C. Michaud, C. J. L. Murray and M. Naghavi (2014), **vol. 134**, no.6, pp. 1527–1534, ISSN 0022-202X, 1523-1747, doi:10.1038/jid.2013.446.
https://www.jidonline.org/article/S0022-202X(15)36827-5/abstract

IKNL (2018), Incidentie huid.
https://www.cijfersoverkanker.nl/selecties/incidentie_huid/img568b9af14d9e9

Kittler, H., H. Pehamberger, K. Wolff and M. Binder (2002), Diagnostic accuracy of dermoscopy, **vol. 3**, no.3, pp. 159–165, ISSN 1470-2045, doi:10.1016/S1470-2045(02)00679-4.
http://www.sciencedirect.com/science/article/pii/S1470204502006794

Maglogiannis, I. and C. N. Doukas (2009), Overview of Advanced Computer Vision Systems for Skin Lesions Characterization, **vol. 13**, no.5, pp. 721–733, ISSN 1089-7771, doi:10.1109/TITB.2009.2017529.

Masood, A. and A. Al-Jumaily (2013), Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms, *International Journal of Biomedical Imaging*, **vol. 2013**, doi:10.1155/2013/323268.

Nasr-Esfahani, E., S. Rafiei, M. H. Jafari, N. Karimi, J. S. Wrobel, S. M. R. Soroushmehr, S. Samavi and K. Najarian (2017), Dense Fully Convolutional Network for Skin Lesion Segmentation, *arXiv:1712.10207 [cs]*.
http://arxiv.org/abs/1712.10207

Organization, W. H. (2016), *Global report on psoriasis*, World Health Organization.

Pal, A., A. Chaturvedi, U. Garain, A. Chandra and R. Chatterjee (2016), Severity grading of psoriatic plaques using deep CNN based multi-task learning, in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 1478–1483, doi:10.1109/ICPR.2016.7899846.

Ronneberger, O., P. Fischer and T. Brox (2015), U-Net: Convolutional Networks for Biomedical Image Segmentation, *arXiv:1505.04597 [cs]*.
http://arxiv.org/abs/1505.04597

Sander, H. M., L. F. Morris, C. M. Phillips, P. E. Harrison and A. Menter (1993), **vol. 28**, no.3, pp. 422–425, ISSN 0190-9622, doi:10.1016/0190-9622(93)70062-X.
http://www.sciencedirect.com/science/article/pii/019096229370062X

Shelhamer, E., J. Long and T. Darrell (2017), Fully Convolutional Networks for Semantic Segmentation, **vol. 39**, no.4, pp. 640–651, ISSN 0162-8828, doi:10.1109/TPAMI.2016.2572683.

Stern, R. S., T. Nijsten, S. R. Feldman, D. J. Margolis and T. Rolstad (2004), **vol. 9**, no.2, pp. 136–139, ISSN 1087-0024, doi:10.1046/j.1087-0024.2003.09102.x.
http://www.sciencedirect.com/science/article/pii/S0022202X15530005

The International Skin Imaging Collaboration: Melanoma Project (2019), ISIC Archive.
https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/isicArchive

The Univeristy of Waterloo (2019), Skin Cancer Detection.
https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection

Tsao, H., J. M. Olazagasti, K. M. Cordoro, J. D. Brewer, S. C. Taylor, J. S. Bordeaux, M.-M. Chren, A. J. Sober, C. Tegeler, R. Bhushan and W. S. Begolka (2015), Early detection of melanoma: Reviewing the ABCDEs, **vol. 72**, no.4, pp. 717–723, ISSN 0190-9622, doi:10.1016/j.jaad.2015.01.025.
http://www.sciencedirect.com/science/article/pii/S0190962215000900

Vestergaard, M. E., P. Macaskill, P. E. Holt and S. W. Menzies (2008), Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting, **vol. 159**, no.3, pp. 669–676, ISSN 1365-2133, doi:10.1111/j.1365-2133.2008.08713.x.
https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2133.

2008.08713.x

de Vries, E., M. Boniol, J. F. Doré and J. W. W. Coebergh (2004), Lower incidence rates but thicker melanomas in Eastern Europe before 1992: a comparison with Western Europe, **vol. 40**, no.7, pp. 1045–1052, ISSN 0959-8049, doi:10.1016/j.ejca.2003.12.021.
http:
//www.sciencedirect.com/science/article/pii/S0959804904000711

Westerhoff, K., W. H. Mccarthy and S. W. Menzies (2000), Increase in the sensitivity for melanoma diagnosis by primary care physicians using skin surface microscopy, **vol. 143**, no.5, pp. 1016–1020, ISSN 1365-2133, doi:10.1046/j.1365-2133.2000.03836.x.
https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2133.2000.03836.x

Yu, F. and V. Koltun (2015), Multi-Scale Context Aggregation by Dilated Convolutions, *arXiv:1511.07122 [cs]*.
http://arxiv.org/abs/1511.07122

Yuan, Y., M. Chao and Y. Lo (2017), Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance, **vol. 36**, no.9, pp. 1876–1886, ISSN 0278-0062, doi:10.1109/TMI.2017.2695227.

Zhang, X., J. Zou, K. He and J. Sun (2016), **vol. 38**, no.10, pp. 1943–1955, ISSN 1939-3539, doi:10.1109/TPAMI.2015.2502579, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Zhao, S., B. Xie, Y. Li, X. Zhao, Y. Kuang, J. Su, X. He, X. Wu, W. Fan, K. Huang, J. Su, Y. Peng, A. A. Navarini, W. Huang and X. Chen (2020), **vol. 34**, no.3, pp. 518–524, ISSN 1468-3083, doi:10.1111/jdv.15965.
https://onlinelibrary.wiley.com/doi/abs/10.1111/jdv.15965