Anette Karhu

Deep Semantic Segmentation for Skin Cancer Detection from Hyperspectral Images

Master's Thesis in Information Technology

December 4, 2020

University of Jyväskylä

Faculty of Information Technology

Author: Anette Karhu

Contact information: anette.n.e.karhu@student.jyu.fi

Supervisors: Ilkka Pölönen, and Sami Äyrämö

Title: Deep Semantic Segmentation for Skin Cancer Detection from Hyperspectral Images

Työn nimi: Hyperspektrikuvien semanttinen segmentointi ihosyövän tunnistamisen apuna

Project: Master's Thesis

Study line: Applied mathematics and computational sciences

Page count: 85+0

Abstract: As skin cancer types are a growing concern worldwide, a new screening tool combined with automation may help the clinicians in clinical examinations of lesions. A novel hyperspectral imager prototype has been noted to be a promising non-invasive tool in screening of lesions. Deep learning, especially semantic segmentation models, have brought successful results in other biomedical imaging tasks. Therefore, semantic segmentation could be used to automate the results from the hyperspectral images of lesions. In this thesis we used a novel hyperspectral image dataset of lesions that contained 61 images. The dataset contained 120 different wavebands from the spectral range of $450-850$ nm with dimensions of 1920×1200 pixels. We implemented two different semantic segmentation models and compared their performance with the novel hyperspectral image data. The models were compared by their ability to segmentate the images and by their ability to classify lesion types from the images. From the implemented models, the combination of ResNet and Unet model architecture (ResNet-Unet) was able to segmentate the images more accurately with f1-score of 92.38 %, whereas the implemented Unet model gained f1-score of 92.17 %. In addition, the ResNet-Unet model classified the lesion types more accurately, and contained only one false negative result in melanoma classification, when the Unet model contained two false negatives in melanoma classification. This study was able to repeat the results of a previous study, where the segmentation model using hyperspectral image data was able to classify melanoma slightly more accurately than the clinicians in a previous study were.

Keywords: biomedical image segmentation, deep learning, hyperspectral imaging, skin cancer, melanoma.

Suomenkielinen tiivistelmä:

Ihosyöpä on maailmanlaajuisesti kasvava ongelma. Sen vuoksi ihosyöpien tunnistamisen avuksi olisi tarpeellista saada uudenlainen diagnostiikkatyökalu terveydenhuollon ammattilaisille. Uusi hyperspektrikuvantamisen prototyyppi on aiemmissa tutkimuksissa osoittautunut lupaavaksi menetelmäksi etenkin ihosyöpätyyppien tunnistamisen tuloksissa. Syväoppiminen, varsinkin semanttinen segmentointi on tuottanut hyviä tuloksia muissa lääketieteellisen kuvantamisen tapauksissa. Segmentointi voisi auttaa myös automatisoimaan luomityyppien tunnistusta hyperspektrikuvista. Tässä työssä käytettiin uutta hyperspektrikuvadataa, joka koostui 61 leesiokuvasta. Data sisälsi yhteensä 120 eri aallonpituutta, alueilta $450-850$ nm ja kuvien dimensiot olivat 1920×1200 pikseliä. Tässä työssä implementoitiin ja vertailtiin kahta eri semanttisen segmentoinnin mallia, käyttäen malleissa uutta hyperspektridataa. Vertailussa tarkasteltiin mallien kykyä segmentoida luomikuvia sekä niiden kykyä tunnistaa luomityypit hyperspektrikuvadatasta. Näistä implementoiduista malleista toinen, kombinaatio ResNet ja Unet arkkitehtuureista (ResNet-Unet), oli parempi molemmissa tehtävissä. Se tuotti kokonaissegmentoinnista f1-metriikalla 92.38 % tarkkuuden, kun implementoitu Unet malli tuotti f1-metriikalla 92.17 % tarkkuuden. ResNet-Unet malli myös tunnisti luomityypit paremmin ja tuotti melanooman tunnistuksessa vain yhden väärän negatiivisen tuloksen, kun Unet malli ennusti kaksi väärää negatiivista tulosta melanoomalle. Kaiken kaikkiaan tässä tutkimuksessa saavutettiin sama tulos kuin aiemmassa tutkimuksessa, eli segmentointimallit pystyivät tunnistamaan melanoomaa hieman tarkemmin kuin mitä aiempi kliininen tutkimus pystyi.

Avainsanat: lääketieteellisten kuvien segmentointi, syväoppiminen, hyperspektrikuvantaminen, ihosyöpä, melanooma.

Notations and abbreviations

$\mathbf{a}$	Activation of artificial neuron
$\hat{a}$	Normalized activations
$\mathbf{b}$	Neuron bias
B	Input batch
$\bar{B}$	Mean of batch
C	Cost function
δ	Training error
f	Activation function
η	Learning rate
G	Residual connection
h	Batch normalization
θ	Models parameters
$\mathbf{I}$	Radiance data
$\mathbf{I}_o$	White reference data
$\mathbf{K}$	Filter of convolution
L	Loss function
$\mathbf{R}$	Reflectance
$\mathbf{s}$	Weighted input
$\mathbf{S}$	Feature map of convolution
V_B	Variance of batch
$\mathbf{w}$	Neuron weights
W	Nonlinear mapping of a layer
$\mathbf{x}$	Input data
$\mathbf{X}$	Input feature map
y	Ground truth
$\hat{y}$	Output predictions
$\mathbf{z}$	Linear activations

AI	Artificial Intelligence
ANN	Artificial Neural Network
BCC	Basal Cell Carcinoma
BN	Benign Nevi
CART	Classification of Regression Tree
CCD	Charged-Coupled Device
CMOS	Complementary Metal Oxide Semiconductor
CNN	Convolutional Neural Network
DN	Dysplastic Nevi
DNN	Deep Neural Network
EM	Electron Microscopy
FCN	Fully Convolutional Network
FPI	Fabry-Pérot interferometer
FWHM	Full Width at Half Maximum
GPU	Graphics Processing Unit
HSI	Hyperspectral Imaging
LM	Lentigo Maligna
LMM	Lentigo Maligna Melanoma
MIS	Melanoma In Situ
MM	Malignant Melanoma
MSI	Multispectral imaging
PPV	Positive Predictive Value
RAM	Random Access Memory
ReLU	Rectified Linear Unit
RGB	Red Green and Blue
RNN	Recurrent Neural Network
ROI	Region of Interest
SCC	Squamous Cell Carcinoma
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine

List of Figures

Figure 1. Examples of the four lesion types we focused on this study 5
Figure 2. A simplified example of image acquisition 9
Figure 3. The electromagnetic spectrum and its wavelengths 10
Figure 4. A simple example of Fabry-Pérot interferometer 11
Figure 5. An example of hyperspectral image visualization 13
Figure 6. Presentation of an artificial neuron 16
Figure 7. The presentations of tanh, sigmoid, and ReLu activation functions 18
Figure 8. An example of artificial neural network with one hidden layer 19
Figure 9. A contour plot of a cost function and finding optima by using gradient descent . 20
Figure 10. An example of the convolution operation with padding 27
Figure 11. An example of the up-convolution operation 28
Figure 12. An example of the pooling method 29
Figure 13. Example outputs of image classification, object detection, and semantic segmentation 30
Figure 14. An example of semantic segmentation 30
Figure 15. Examples of lesion types used in the study 36
Figure 16. A false color hyperspectral image and an annotated ground truth image. 38
Figure 17. The architecture of the implemented Unet model 41
Figure 18. The architecture of the implemented ResNet and Unet model combination (ResNet-Unet) 42
Figure 19. Examples of high-quality segmentation results of lesions by the implemented Unet model 50
Figure 20. Examples of low-quality segmentation results of lesions by the implemented Unet model 51
Figure 21. Examples of high-quality segmentation results of lesions by the implemented ResNet-Unet model 53
Figure 22. Examples of low-quality segmentation results of lesions by the implemented ResNet-Unet model 54
Figure 23. The confusion matrices of both models overall lesion classification over the 5-fold cross-validation 56
Figure 24. The overall lesion types predicted for each true lesion type by the implemented Unet model 59
Figure 25. The overall lesion types predicted for each true lesion type by the implemented ResNet-Unet model 60

List of Tables

Table 1. The amount of lesion type examples the used dataset contained. 35

Table 2. Hyperspectral imaging systems specifications used for the image capturing. The more detailed description of the imaging system can be found from Pölönen et al. (2019). 37
Table 3. The summary of the amount of parameters the models contained. 44
Table 4. The risk classes for the lesion types used in this study. Risk class one presents the most severe lesion type and risk class four presents not dangerous lesion type... 47
Table 5. The summary of the segmentation results between the two implemented models. The table shows the mean and standard deviation of the experimental results on both models over the 5-fold cross-validation. 49
Table 6. The summary of the malignant melanoma classification on both models. The table shows the malignant melanoma classification mean and standard deviation results of the 5-fold cross-validation. 57
Table 7. The summary of the lentigo maligna classification on both models. The table shows the lentigo maligna classification mean and standard deviation results of the 5-fold cross-validation. 57
Table 8. The summary of dysplastic nevus classification on both models. The table shows the dysplastic nevus classification mean and standard deviation results of the 5-fold cross-validation. 58
Table 9. The summary of bening nevus classification on both models. The table shows the bening nevus classification mean and standard deviation results of the 5-fold cross-validation. 58
Table 10. Comparison of the implemented ResNet-Unet model and the 2D CNN model (Pölönen et al. 2019) results with all of the lesion types. The dashes in the table indicate that the results were not reported with such metric. 63

Contents

1 INTRODUCTION 1
1.1 Problem statement 2
1.2 Structure of the thesis 3

2 THEORETICAL BACKGROUND 4
2.1 Skin cancer 4
2.1.1 The development of skin cancer 6
2.1.2 The screening of skin cancer 6
2.2 Hyperspectral imaging 8
2.2.1 Radiation 8
2.2.2 Spectral imaging 11
2.2.3 Biomedical hyperspectral imaging 14
2.3 Deep learning 15
2.3.1 Artificial Neural Networks 16
2.3.2 Training and optimizing deep learning models 19
2.3.3 Convolutional neural networks 25
2.4 Semantic segmentation 29
2.4.1 Semantic segmentation in skin cancer detection from hyperspectral images 32

3 MATERIALS AND METHODS 35
3.1 Materials 35
3.2 Methods 38
3.2.1 Experiments 43
3.3 Evaluation 44
3.3.1 Evaluation of semantic segmentation 45
3.3.2 Evaluation of lesion classification 46

4 RESULTS 48
4.1 Semantic segmentation results 48
4.1.1 Semantic segmentation results of the implemented U-net model 48
4.1.2 Semantic segmentation results of the implemented ResNet-Unet model 52
4.2 Lesion classification 55

5 DISCUSSION 61
5.1 Suggestions for further research 65

6 CONCLUSIONS 66

BIBLIOGRAPHY 67

1 Introduction

There is evidence that skin cancer diagnoses have increased worldwide (Jerant et al. 2000). Especially melanoma incidences have been shown to increase rapidly in several countries (Hall et al. 1999; Jerant et al. 2000; Lasithiotakis et al. 2006; Stang et al. 2006). Melanoma is the most dangerous skin cancer type as it has the highest mortality rate (Cummins et al. 2006; Jerant et al. 2000). Early detection is crucial for skin cancer and melanoma detection (Cummins et al. 2006). Unfortunately, the determination of skin cancer with current tools available is challenging, as lesion types can visually resemble each other. Therefore, there is a need for a noninvasive diagnostic tool in clinical usage to help with skin cancer detection. A new tool would help to gain more accurate diagnosis already in clinical examination, and help diagnosing tumour in the early stage. In addition, it is possible that this would also help to decrease the societal costs of skin cancer treatments (Eriksson and Tinghög 2015).

A novel hyperspectral imager combined with deep learning have been reported to be a useful tool in skin cancer detection (Neittaanmäki-Perttu et al. 2013; Pölönen et al. 2019). As a regular camera uses only three wavebands in imaging, a hyperspectral imager can use tens or hundreds of wavebands when capturing an image. Therefore, hyperspectral images can bring more information from lesions and the surrounding tissues. As concluded by Neittaanmäki-Perttu et al. (2015) and Salmivuori et al. (2019) this imaging tool can help to prevent unneeded lesion removals when tumours can be delineated more accurately. Hyperspectral imaging have also been noted to help in detecting skin cancer in the earlier stage. (Neittaanmäki et al. 2017).

Deep learning could be used to automate the lesion segmentation and the lesion classification process from the novel hyperspectral image data. Deep learning models have been adopted to automate several different tasks such as scene understanding (Badrinarayanan et al. 2015), autonomous driving (Sallab et al. 2017), and biomedics (Ronneberger et al. 2015). Convolutional neural networks (CNNs) in deep learning have been able to succeed with remarkable results in several tasks (Badrinarayanan et al. 2015; He et al. 2015a; Krizhevsky et al. 2012). Especially the semantic segmentation of biomedical images with CNNs have brought great results (Ciresan et al. 2012; Ronneberger et al. 2015). Therefore, automating the data processing of hyperspectral images could enable to gain faster and more accurate diagnosis, perhaps already in the clinical examination.

Several deep learning methods have been used to segmentate skin cancer but most of the studies were conducted by using RGB or multispectral images, or their combinations (Gorriz et al. 2017; Yu et al. 2017; Alom et al. 2018). There are less studies that use convolutional neural networks to segmentate malignant melanoma from hyperspectral data (Pölönen et al. 2019). Therefore, motivated by the promising results of using convolutional neural networks in semantic segmentation of lesions, we attempt to find benefits by using two different deep learning architectures to segmentate lesions and to classify lesions from novel hyperspectral image data.

1.1 Problem statement

This study focuses on implementing and comparing two architecturally different semantic segmentation models – the Unet model (Ronneberger et al. 2015) and the ResNet model (He et al. 2015a) combined with the Unet architecture (ResNet-Unet). We evaluate the two models ability to segmentate lesions and classify lesions from the novel hyperspectral image data. A recent study by Pölönen et al. (2019) was able to gain successful results on classifying malignant melanoma from a novel hyperspectral image dataset they collected. However, the models used in the study had some difficulties to accurately segmentate the borders of the lesion images. (Pölönen et al. 2019). Therefore, this study aims to test two different deep learning models and compare the overall semantic segmentation and the classification capability of different lesion types. In addition, we compare the lesion classification accuracy of our implemented models with the results of the study by Pölönen et al. (2019).

The following research questions are answered in this research:

1. Which deep learning architecture gained best results on semantic segmentation from hyperspectral images of lesions?
2. Which deep learning architecture gained best results on classifying different lesion types from hyperspectral images of lesions?
3. Can either of the implemented models improve the classification of different lesion

types from hyperspectral images when compared to the study by Pölönen et al. (2019)?

1.2 Structure of the thesis

The structure of this thesis has been organised in the following way. First, in Chapter 2 the theoretical background of this thesis is introduced. Then, in Chapter 3 the materials and the methods of the research are described. Next, in Chapter 4 the results of the research are explained. In Chapter 5, the findings of the study are discussed in detail and we present the potential future work. Finally, in Chapter 6 we give conclusions for this study.

2 Theoretical background

In this chapter, the theoretical background for the thesis is introduced. This chapter is composed of four main elements: skin cancer, hyperspectral imaging, deep learning and semantic segmentation, which present the main concepts of this study. First, in Section 2.1 a brief overview of skin cancer and its current treatment is covered. Then, the novel imaging method, hyperspectral imaging is presented in Section 2.2. Next, the key aspects of deep learning and artificial neural networks are discussed in detail in Section 2.3. Finally, semantic segmentation, the method used to automate the skin cancer predictions from hyperspectral image data, is introduced in Section 2.4.

2.1 Skin cancer

In this section we will focus on introducing skin cancer. In Section 2.1.1 we will go trough the risks of skin cancer and how it is developed. Finally, in Section 2.1.2 we will introduce the current screening methods of skin lesions.

Incidents of skin cancer have increased during several years (Jerant et al. 2000). Especially incidences of melanoma, the deadliest form of skin cancers, have increased rapidly (Hall et al. 1999; Jerant et al. 2000). The death rate of melanoma has increased, even though the survival rate has improved during the years (Rigel and Carucci 2000). The trend of growing amount of melanoma diagnoses are estimated to continue in future (Siegel et al. 2019).

Skin cancer types are usually presented as melanoma or non-melanoma. Non-melanoma types of skin cancer are usually divided into two groups – basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). Non-melanoma types of skin cancer have higher incident rate than melanoma. (Guy and Ekwueme 2011; Jerant et al. 2000). Non-melanoma tumours do not tend to metastasize, in other words spread to other parts of the body. In fact, they can be treated quite well and they have a low mortality rate. (Guy and Ekwueme 2011; Jerant et al. 2000). Melanoma, on the other hand, has a high mortality rate. Infact, it is the deadliest skin cancer type. (Cummins et al. 2006; Jerant et al. 2000). Melanomas tend to metastasize and become more aggressive over time, and treatment of metastatic melanoma is hard. Due to this reason, melanoma should be detected in its early stage, when the treatment is easier and mortality and treatment costs are lower. (Eriksson and Tinghög 2015; Marghoob et al. 2003; Weinstock 2006).

The four lesion types we focused on this study, vary from melanoma to benign. A sample of each lesion type is visualized in Figure 1. Next we will introduce the four lesion types in more detail.

- **Benign nevi (BN)** is a normal and non-malignant mole, which is a very common lesion type. Few benign lesions may resemble malignant melanoma (Jerant et al. 2000).
- **Dysplastic nevi (DN)** is also a non-malignant but atypical mole. DN have a risk to develop into melanoma which is why their screening is important (Rigel et al. 1989).
- **Lentigo Maligna (LM)**, also known as melanoma in situ (MIS), is a malignant tumour that has not yet spread to other parts of body. LM might develop into malignant melanoma, therefore LM should be excised (Tannous et al. 2000). The treatment costs of LM have shown to be lower than malignant melanomas (Alexandrescu 2009).
- **Malignant melanoma (MM)** is the most aggressive type of skin cancer. MM can develop metastases, which is the reason of the high mortality rate of MM when compared to the other skin cancer types. (Jerant et al. 2000).

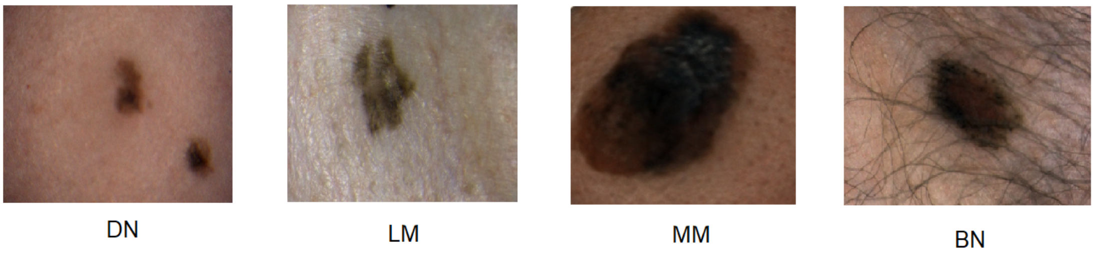

Figure 1. Examples of the four lesion types we focused on this study: Dysplastic nevi (DN), Lentigo Maligna (LM), Malignant melanoma (MM), and Benign nevi (BN). Images are from the dataset that was used in this study.

2.1.1 The development of skin cancer

As most cancers, skin cancer starts with small changes in the body, also called as precancerous stages. Skin cancer may develop into atypical, non-normal moles or unusual skin growths (Jerant et al. 2000; Tsao et al. 2004). These atypical lesions and dysplasias should be followed carefully, as malignant lesions tend to change during time (Tsao et al. 2004). Early detection of melanoma lowers the mortality rate significantly (Rigel and Carucci 2000). However, the detection might be hard, as malignant lesions can visually resemble typical moles in their early-stage (Jerant et al. 2000; Rigel and Carucci 2000). To improve early detection of melanoma, novel and enhanced imaging tools for clinical observations could help to delineate skin cancer from healthy tissues.

There are several risk factors how skin cancer may develop. Typical risks are age (Siegel et al. 2019), a large amount of nevi (Gandini et al. 2005a), and previous history of sun exposure (Gandini et al. 2005b). People with lighter skin tone usually have been reported with most of the skin cancer incidents. Nevertheless, people with color have been noted to have a very high mortality rate with skin cancer when compared to people with lighter skin color. (Gloster Jr and Neal 2006). Furthermore, men are more likely to develop aggressive skin tumours (Jerant et al. 2000). If skin cancer has been diagnosed in a family, it has been noted to increase the risk that a family member may develop inherited melanoma (Greene et al. 1985). In conclusion, skin cancer can develop during time treacherously to anyone. It seems that the best methods to detect skin tumours in early-stage are screening of lesions and raising public awareness of the dangers with exposure to the sun. (Jerant et al. 2000; Rigel and Carucci 2000).

2.1.2 The screening of skin cancer

A key aspect of detecting skin cancer is the screening process. Screening is a systematical process where the patient's lesions are reviewed by a healthcare expert. (Jerant et al. 2000). The most common screening method for recognising melanoma is to follow the ABCD guideline, which stands for asymmetry, border irregularity, color variegation, and diameter. (Friedman et al. 1985). The detection of early-stage melanoma can be very difficult, as the lesion types may resemble each other (Jerant et al. 2000; Rigel and Carucci 2000; Tsao et al. 2004) For example, a study by Heal et al. (2008) inspected the diagnoses of different skin cancers by skin specialists and general practitioners, and compared their diagnoses to histopathologically verified diagnoses. They reported recall of 33.8 %, and precision of 33.3 % in melanoma detection, whereas non-melanoma diagnoses were more likely to be classified correctly. It seems that accuracy of melanoma diagnosis in clinical examinations relies strongly on the observer's expertise, thus education and diagnostic tools can improve the detection accuracy (Argenziano et al. 2012; Heal et al. 2008; Offidani et al. 2002).

The diagnoses of lesions are verified with histopathological examination. If a clinician detects an atypical lesion in the patient's skin, then a biopsy is taken from the lesion. The histopathological analysis, also known as microscopic examination of the tissue is performed to study the biopsy. This analysis returns the most dangerous lesion type diagnosis of the lesion. Histopathological analysis is currently the best method to gain reliable disease classification information from a biopsy of a lesion (Rigel and Carucci 2000; Tsao et al. 2004). Sometimes re-excisions of skin tumours are needed after verified results from histopathological examination (Tsao et al. 2004). Although, the current method has some downsides. For example, if a person has multiple atypical lesions the excision of all lesions cannot be performed (Rigel and Carucci 2000). Unnecessary excisions should be avoided as they might develop infections for patients. Moreover, the method also depends on the pathologists expertise, results need time to be conducted, and it is quite expensive. (Liu et al. 2011). Certainly, a new non-invasive clinical tool is desired to improve and speed up the clinical detection of skin cancer, as the incidences of skin cancer continue to increase (Liu et al. 2011; Rigel and Carucci 2000).

Currently, there are a few imaging tools on the market to help healthcare experts to detect skin cancer in clinical examinations. Today, one of the most common imaging tool used in clinical examinations is dermoscopy (Braun et al. 2005). As Vestergaard et al. (2008) reviewed, the usage of dermoscopy in clinical examinations can improve the diagnostic accuracy of melanoma, when compared to clinical examinations without any diagnostic tools. According to Braun et al. (2005), the accuracy of dermoscopy diagnosis is lower with inexperienced practitioners, therefore an automated diagnosis tool could be helpful in clinical observations.

Fink and Haenssle (2017) point out that even though dermoscopy can improve the diagnostic accuracy, the diagnoses are always verified histopathologically. Therefore, it has not decreased excisions of lesions or replaced histopathological examinations (Fink and Haenssle 2017). Moreover, dermoscope combined with deep learning automation, has not been able to outperform experienced dermatologists (Esteva et al. 2017). Johansen et al. (2020) argued that the dermoscopic systems performance may not be possible to improve. Therefore, new imaging methods and tools could be the key to improve diagnostic accuracy, for example by using hyperspectral imaging combined with deep learning (Johansen et al. 2020). Therefore, in this study we use novel hyperspectral image data of lesions. In conclusion, hyperspectral imaging combined with automated prediction could be a useful tool in clinical examinations of lesions (Neittaanmäki-Perttu et al. 2013; Salmivuori et al. 2019).

2.2 Hyperspectral imaging

The previous section introduced the types of skin cancer, especially melanoma type of skin cancer. We also presented the challenges in the current skin cancer screening tools. Therefore, we will now discuss about a novel imaging method that could improve current lesion screening. Hyperspectral imaging has been developed in order to gain more information of the surroundings by using more spectral bands in imaging. This section describes the fundamentals of hyperspectral imaging. First, the electromagnetic spectrum and the methods to separate the spectral bands are presented. Then, the spectral imaging is introduced in more detail. Lastly, we will focus on biomedical hyperspectral imaging.

2.2.1 Radiation

The imaging process is based on capturing the electromagnetic radiation that has been reflected of the objects being imaged. In regular cameras the visible spectrum has been the most used portion of spectra, whereas the sensors of hyperspectral imagers and multispectral imagers can acquire other portions of the electromagnetic spectrum in addition to visible light. This way non-visible wavelengths and visible wavelength ranges can be observed and more information can be gained from the imaged object. (Chang 2007, Chapter 2). The basic principle of capturing an image from an object and its surroundings is presented in Figure 2.

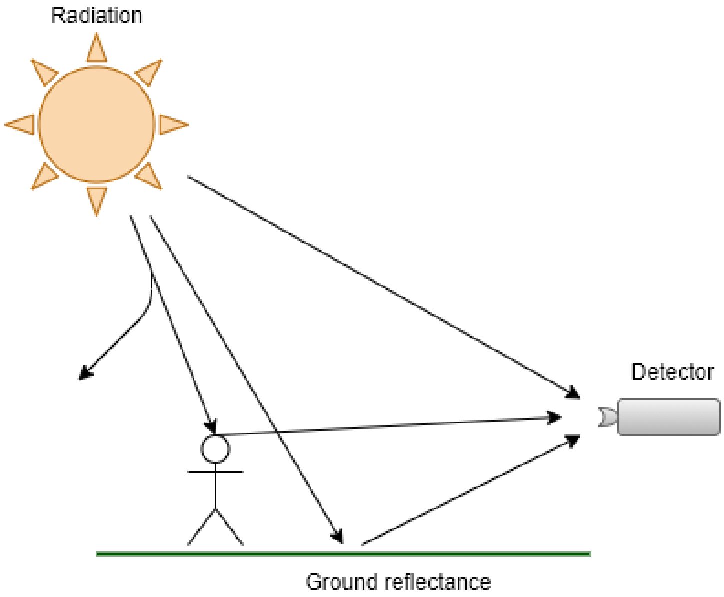

Figure 2. A simplified example of image acquisition. The detector in imaging devices capture the radiation from the imaged objects and its surroundings. Some of the radiation is usually scattered and absorbed by different materials and matter. A portion of the radiance is reflected of the objects, which the detector records.

The electromagnetic spectrum represents radiation and it can be divided into the following main regions: gamma rays, x-rays, ultraviolet, visible spectrum, infrared, microwaves, and radio waves. A human can only see the visible spectrum, approximately from 400 – 700 nanometers. Therefore, the other wavelengths are referred as non-visible ranges, and they can be observed with detectors of imagers. The electromagnetic spectrum is divided into aforementioned separate ranges by the length of the wavelengths and by the difference of interaction with matter. (Stuart 2004). The electromagnetic spectrum is visualized in Figure 3.

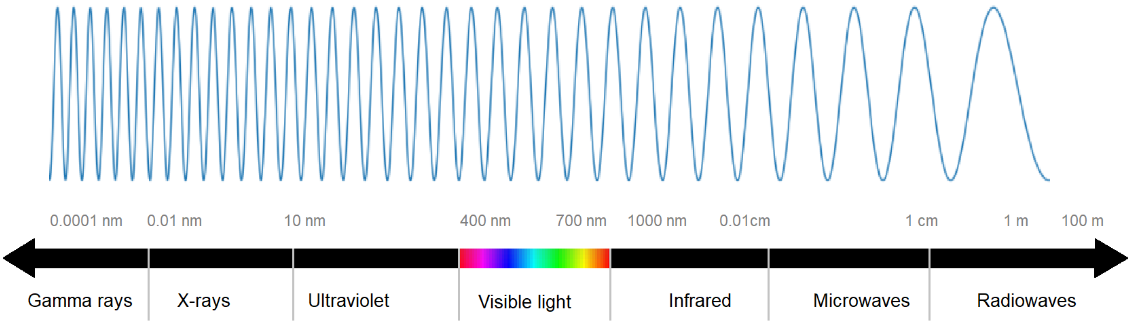

Figure 3. The electromagnetic spectrum and its wavelengths. On the left side of the image the spectral regions with shorter wavelengths and higher energy levels are presented. On the right side the regions with longer wavelengths and lower energy levels are shown.

Using visible light and non-visible wavelengths in imaging process can help to identify and gain more information from the imaged objects and the surroundings. This allows to observe spectral signatures of objects in a larger range of wavelengths. Spectral signatures describe the amount of radiation reflected from an object over a spectral range. (Jones and Vaughan 2010). Each material has somewhat unique spectral characteristics but these spectral signatures may have some variation, for example over time and over space (Chang 2007; Jones and Vaughan 2010, Chapter 2). These structures and characteristics of matter, and their interaction with electromagnetic spectrum are studied in the field of spectroscopy (Wolfe 1997).

The separate wavelengths of electromagnetic spectrum can be measured with, for example an interferometer or a triangular prism. A triangular prism is a method to disperse light, but the measurements can be acquired only by one line at a time. (Garini et al. 2006). The detectors of interferometers allow the whole spectrum of wavelengths to be used concurrently (Harvey 2011, Chapter 10). Optical interferometry uses interference patterns that are processed to acquire the specific spectrum, for example by using inverse Fourier transform (Hariharan 2010; Chang 2007, Chapter 2).

2.2.2 Spectral imaging

Having introduced the basics of electromagnetic spectrum and radiation, we will now discuss the spectral imaging in more detail. Spectral imaging combines spectroscopy, the study of material and radiation interaction, with imaging. Spectral imaging provides spatial and spectral information from the imaged object. The spectral range that is most commonly recorded in spectral imaging combines one or more wavelengths from the following regions: ultraviolet, visible light, near-infrared, and mid-infrared. (Garini et al. 2006; Chang 2007, Chapter 2).

There are several methods to record spectral information in spectral imaging. One method is to use previously introduced interferometers, such as Fabry-Pérot interferometer, which can be seen in Figure 4. Fabry-Pérot interferometer has two partly reflecting mirrors separated by an air gap, and followed by a lens before the detector. (Vaughan 1989). The air gap in the Fabry-Pérot interferometers allow to tune the observed wavelengths. This allows to change the observed wavelengths easily by only adjusting the parameters. (Saari et al. 2010).

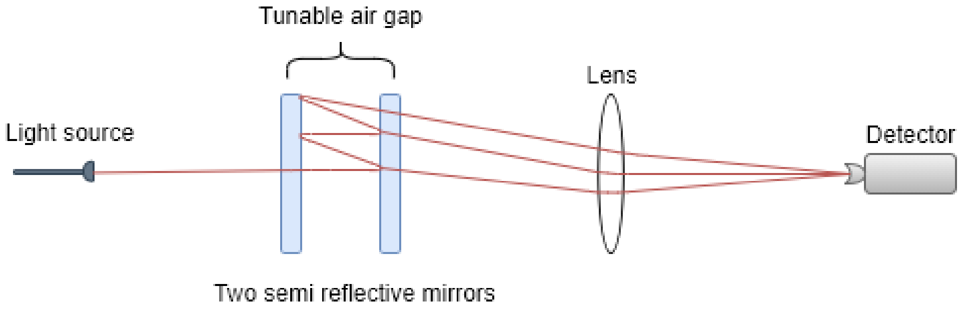

Figure 4. A simple example of Fabry-Pérot interferometer. The two partly reflective mirrors have a tunable air gap in between, followed by a lens. The last element in the interferometer is a detector. Radiation into the interferometer is provided by using an external light source.

The image acquisition in spectral imaging is performed by using a detector. Detectors transfer the radiation into digital numbers. (Chang 2007, Chapter 2). The sensors of digital cameras change the radiation into electrons (Garini et al. 2006). There are several sensors available, but few of the most common sensors are the complementary metal–oxide–semiconductor (CMOS) and the charged-coupled device (CCD). (Lu and Fei 2014). The quality of spectral images are commonly presented with spectral resolution, which implicates the imagers capability to measure and distinguish spectral features. Full width at half maximum (FWHM) presents the width of the spectrum being observed. (Sun 2010, Chapter 1). Together these metrics provide information of the accuracy and quality of the spectral images. (Lu and Fei 2014).

Both multispectral imaging and hyperspectral imaging are subcategories of spectral imaging. The difference between these imaging methods are that multispectral imaging (MSI) usually acquires the images by using less than ten separate wavebands. The pixels of multispectral images do not form a continuous spectrum from the object being imaged. (Chang 2007, Chapter 2). In contrast to MSI, Hyperspectral imaging (HSI) can capture tens and even hundreds of narrow and contiguous wavebands. (Chang 2007, Chapter 2). Therefore, hyperspectral images are often referred to contain a continuous spectral curve from the imaged object in each pixel of the image (Johansen et al. 2020). The continuous spectrum from the imaged target enables HSI system to record more information, whereas MSI may lack some important information (Lu and Fei 2014).

Hyperspectral images are usually presented as a data cube, which is demonstrated on the left side of Figure 5. In the data cube the height and the length of the cube represent the dimensions of an image, and the width shows the number of wavelength channels used. Each pixel on the data cube represents the spectrum of the pixel, whereas, each image layer shows the image in a specific wavelength. Another method to present the spectral data is to plot the pixel-wise spectral curve, which is seen on the right side of the figure.

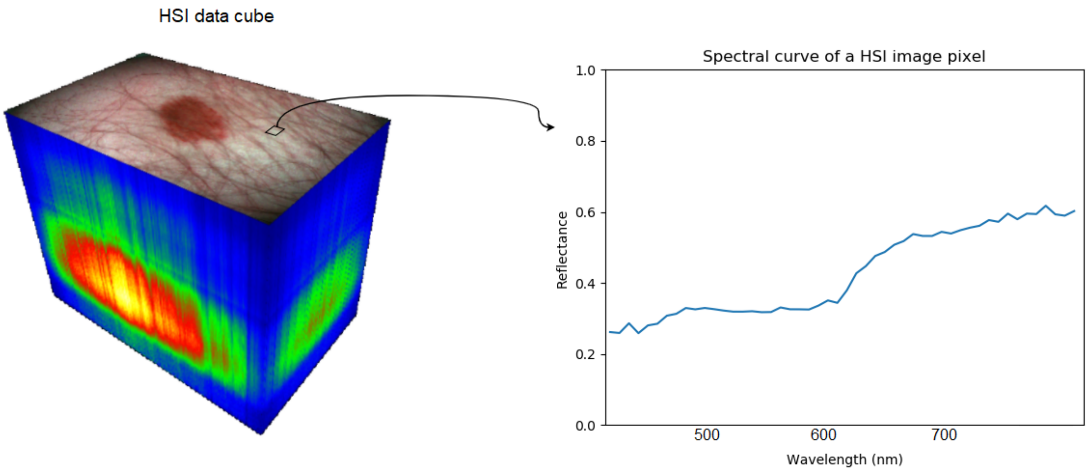

Figure 5. On the left side we can see the hyperspectral image as a data cube presentation. (Reproduced from Boggs (2014)). On top of the image is the RGB presentation of a lesion, and the channels of the hyperspectral image are shown as depth in the image. On the right side a spectrum of a HSI image single pixel is visualized.

Even though hyperspectral imaging has many advantages, there are also some downsides with the imaging method. For example, the HSI system can be quite expensive (Saari et al. 2010). Also, the size of hyperspectral image can be quite large, as every pixel can contain hundreds of spectral bands of information. This affects to the processing time of the image (Garini et al. 2006). Moreover, the HSI system needs to be calibrated before capturing images from a specific target. This ensures that the spectral quality of the image meets the case specific requirements. (Sun 2010, Chapter 1). The produced HSI data also needs to be processed and analysed in order to interpret the images. (Lu and Fei 2014).

The preprocessing of HSI data include several steps, such as normalizing the data, calibrating the observed wavelengths and reducing the noise effects. (Sun 2010, Chapter 2). For example, the hyperspectral imager can present the image in raw data format as digital numbers that can be converted to radiance. In addition, the data can be converted from radiance to reflectance. This operation corrects the spectrum of the pixels in an image to present only the spectrum of the imaged surface material, and it can minimize the noise effects from external light source. (Sun 2010, Chapter 2). The mathematical equation to gain reflectance from radiance data cubes and from the white reference diffuse reflectance data cubes is the following:

$$\mathbf{R} = \frac{\mathbf{I}}{\mathbf{I}_o} \tag{2.1}$$

where $\mathbf{R}$ denotes the reflectance, $\mathbf{I}$ denotes the radiance data cube, and $\mathbf{I}_o$ denotes the white reference data cube (Pölönen et al. 2019). The preprocessed HSI data can then be further analyzed, for example by using feature extraction methods to downsample the image dimensions. Reducing the image size can help to reduce nonrelevant information from the data, but also it helps to process the data faster. (Lu and Fei 2014).

2.2.3 Biomedical hyperspectral imaging

Hyperspectral imaging has been successfully used in several fields, such as in remote sensing (Adam et al. 2010; Govender et al. 2007), in food safety (Feng and Sun 2012), in crime scene detection (Schuler et al. 2012), and in biomedical imaging (Carrasco et al. 2003; Neittaanmäki et al. 2017; Salmivuori et al. 2019). Especially, in biomedical imaging the HSI has shown to be a promising imaging tool, as it enables to diagnose several illnesses beyond the visible sight and without the need of excisions (Johansen et al. 2020). In biomedical imaging HSI systems are usually calibrated to record specific wavelengths, such as ultraviolet, visible light, and near-infrared regions, which can be selected case specifically according to the optical properties of the imaged biological tissue (Lu and Fei 2014). The usage of non-visible wavelengths enables to identify tissues by their spectral signatures. In addition, the non-visible wavelengths can penetrate slightly further from the surface of the skin, and therefore bring more information from the tissues. (Lu and Fei 2014; Salzer et al. 2000).

Recently, there have been several studies focusing on improving skin cancer detection and delineating the tumour borders by using a novel hyperspectral imaging system. (Neittaanmäki-Perttu et al. 2013; Neittaanmäki-Perttu et al. 2015; Zheludev et al. 2015). A study by Neittaanmäki-Perttu et al. (2013) reported that by using a novel HSI system they were able to detect skin field cancerisation more specifically than with regular clinical observation methods. In another study, Neittaanmäki-Perttu et al. (2015) studied skin cancer types of

LM and lentigo maligna melanoma (LMM) by using a novel HSI system in order to detect these tumour borders more accurately. The study found that the novel HSI was able to delineate the lesions more specifically when compared to regular clinical observations. In addition they hypothesised that the novel HSI system could spare the amount of excised tissue, avoid re-excisions of lesions, and help clinicians in the skin cancer diagnosis process (Neittaanmäki-Perttu et al. 2015). Zheludev et al. (2015) demonstrated the usage of supervised machine learning method, classification and regression tree (CART), in order to detect skin cancer borders from hyperspectral images of lesions. Zheludev et al. (2015) found that the supervised machine learning method was able to detect areas of skin tumours efficiently from the novel hyperspectral lesion data, but further development is needed.

Hyperspectral imaging system seems to be a potential tool in biomedical imaging, especially in skin cancer detection and tumour delineation. Moreover, hyperspectral images of skin cancer interpretation and results could be further improved and automated by using unsupervised methods, such as deep learning. The interpretation capabilities of deep learning methods with novel HSI data of skin lesions have not yet been studied in great detail. (Johansen et al. 2020).

2.3 Deep learning

In the previous section we focused on hyperspectral imaging, and especially its usage in biomedical imaging tasks. This section describes the basics of deep learning. First, the structure of artificial neural networks (ANNs) is described. Next, the training process of ANNs is presented. Finally, the state of the art in deep learning, convolutional neural networks are introduced.

Deep learning methods can be used to automate tasks and they can be used in very complex problems. Deep learning methods learn by themselves from the data they are provided. (Goodfellow et al. 2016). For example, deep learning has been used to beat the professional players in the game of Go (Silver et al. 2016). Furthermore, deep learning has been widely adopted in different fields, for instance it has been used in image classification (Krizhevsky et al. 2012), in road area segmentation (Meyer et al. 2018; Oliveira et al. 2016), in text recognition (Jaderberg et al. 2014; Kai Wang et al. 2011), and in biomedical image analysis (Ciresan et al. 2012; Ronneberger et al. 2015).

2.3.1 Artificial Neural Networks

Artificial neural networks (ANNs) are the foundation of deep learning. ANNs have been inspired by the neuron and the brain study conducted by McCulloch and Pitts (1943), but also by the perceptron model developed by Rosenblatt (1957). These findings and presentations are used today in the building blocks of ANNs. From these studies the mathematical presentation of artificial neurons in deep learning have been adopted from. Although, one must bear in mind that these presentations do not present the biological neurons, which are more complex structures. (Goodfellow et al. 2016).

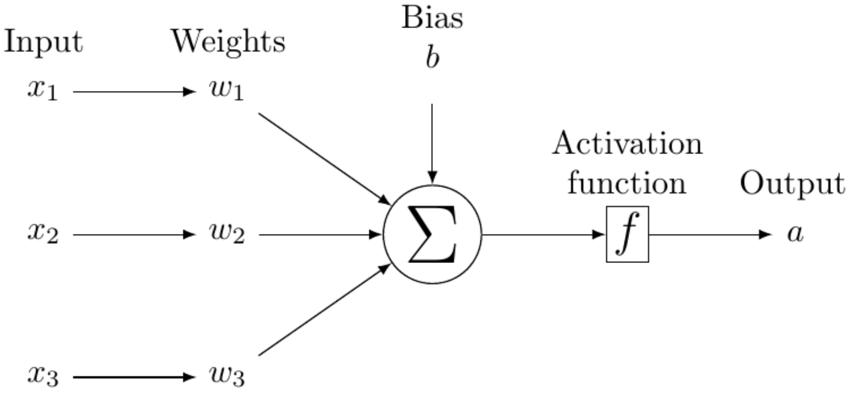

Figure 6. Presentation of an artificial neuron. The input data x_1,x_2, and x_3 are first multiplied with weights w_1, w_2, and w_3. The weighted inputs are then added together with a bias term b. Finally, the linear activations are passed to an activation function f that present the output a.

Artificial neural networks contain several layers. Each layer in the network consists of several artificial neurons. A single artificial neuron is presented in Figure 6. Artificial neurons are connected to each other in a layer-wise manner, and these connections are also known as weights. The mathematical equation for an artificial neuron is the following:

$$a = f\left(\sum_i w_i x_i + b\right) = f(\mathbf{w}^T \mathbf{x} + \mathbf{b}) \tag{2.2}$$

where a is the output of the neuron, $\mathbf{x}$ are the inputs, $\mathbf{w}$ denotes the weights, $\mathbf{b}$ is the bias, and f is the activation function. The sum of inputs x_i and weights w_i shows the importance of a connection. The learnable bias parameter is used to shift the prediction of the network. (Bishop 2006, pp. 227-229).

To nonlinearize the neural networks we use activation functions. Some of the most traditional activation functions are tanh, sigmoid, softmax, and rectified linear unit (ReLu). Few of the activation functions are visualized in Figure 7. The sigmoid activation function transfers the output probabilities between zero and one by:

$$f(\mathbf{z}) = \frac{1}{1 + e^{-\mathbf{z}}}, \tag{2.3}$$

where $\mathbf{z} = \mathbf{w}^T \mathbf{x} + \mathbf{b}$. Here $\mathbf{x}$ is the input feature vector, $\mathbf{w}$ denotes the weights, and $\mathbf{b}$ denotes the bias. The sigmoid is usually used with binary classification tasks. Another binary classification activation function is the tanh activation function:

$$f(\mathbf{z}) = \tanh(\mathbf{z}), \tag{2.4}$$

where $\mathbf{z}$ presents the linear activation. Tanh transfers the output values of a model between values $[-1, 1]$. While tanh and sigmoid activation functions have been traditionally used widely, it was noted that they lack of improving the weights over time. This problem is known as saturation and vanishing gradient problem. (Goodfellow et al. 2016, pp. 191-192). To meet the challenges, ReLu was introduced, which does not have the problem of vanishing gradients (Goodfellow et al. 2016, pp. 187-191). This activation function has the following form:

$$f(\mathbf{z}) = \max(0, \mathbf{z}). \tag{2.5}$$

The advantages of the ReLU activation function is the speed of calculation and the fast convergence. (Goodfellow et al. 2016, pp. 187-191). Although, at times the neurons of a network may stop learning, a problem that is known as the dying ReLU. There has been progress in order to fix this problem by having functions differentiable at all points, such as the leaky ReLU. (Goodfellow et al. 2016, pp. 187-191). Finally, we introduce the activation function which is widely adopted in multi-class predictions, the softmax function. The softmax has the following form:

$$f(\mathbf{z})_i = \frac{exp(z_i)}{\sum_{j=1}^{k} exp(z_j)} \text{ for } i = 1, \ldots, k, \tag{2.6}$$

where $\mathbf{z} \in \mathbb{R}^k$ presents the linear activation, i denotes each element of the linear activation, and k refers to the amount of output classes. The softmax is applied to produce outputs between values zero and one, which all together sum up to one. (Bishop 2006).

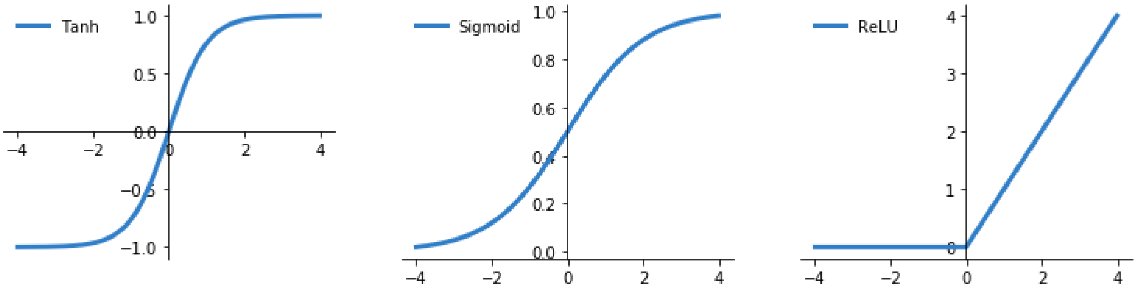

Figure 7. The presentations of tanh, sigmoid, and ReLu activation functions.

The artificial neural network consists of several neurons that are connected to each other, an example can be seen in Figure 8. Together these neurons build up a neural network with multiple layers. These layers consist of an input layer, hidden layers, and an output layer. (Bishop 2006, pp. 227-229). The information in the ANN can go from layer to another in several ways, for example in feedforward or in recurrent manner. The feedforward networks operate and forward information from layer to another only in one direction. (Goodfellow et al. 2016, pp. 164-167). Recurrent neural networks (RNNs), on the other hand, can have cycles or loops in their network structure. The information of loops can be saved, which may help in future predictions. (Goodfellow et al. 2016, pp. 372-376). In addition, the information in the networks can flow straight forward or by using skip connections in a network, which allows to pass information from upper level to lower level by skipping some amount of layers in between. (He et al. 2015a).

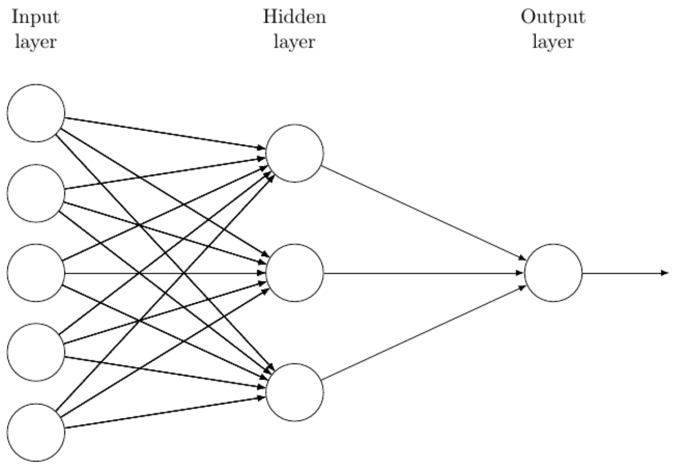

Figure 8. Artificial neural network with one hidden layer

There are several ways that the artificial neural networks can be trained. The usual way of the ANNs training is called supervised learning. In supervised learning the network is trained by using training data together with labelled data. The ANNs can also be trained in unsupervised manner. In unsupervised learning the model learns to predict from training data without the labelled data. When a model is trained the network prediction performance is usually tested with new unseen data, called the test data. (Goodfellow et al. 2016).

Deep learning models are usually deeper presentations of ANNs, also known as deep neural networks (DNNs). The usual idea in deep learning is that the deeper the model, the better the performance (Simonyan and Zisserman 2014). Although, when the depth of a network is increased the more they suffer from the curse of dimensionality. This is straight forward, as the more parameters a network contains the more configurations there are, which increases the difficulty of optimizing a network. (Goodfellow et al. 2016, pp. 152-154). Next we will discuss the training of deep learning models and their optimization in more detail.

2.3.2 Training and optimizing deep learning models

Training deep learning models is a challenging task, as there are many learnable and adjustable parameters in the network that need to be optimized to gain reliable results. We will now focus on the building blocks of training deep learning models – parameter initialization, cost function, backpropagation, and optimization methods.

The training of a model begins by initializing the parameters of a network in a random manner. This enables to distinguish the parameters by eliminating the symmetry between the parameters. Initialization of parameters is essential, as they affect on the networks convergence. Moreover, a bad choice of parameters can lead into exploding gradient problem or into vanishing gradient problem. (Goodfellow et al. 2016, pp. 296-302). There are several methods to initialize the parameters, for example by using the Glorot uniform initialization (Glorot and Bengio 2010), or by using the HE normal initialization (He et al. 2015b).

An important aspect in training a deep learning model is optimization. As we have discussed earlier, the supervised models are trained by using training data and ground truth data. As the model makes predictions from the training data, the output of the model is then compared with the desired output. From here the cost function of a model can be calculated. The model then tries to minimize this cost function $C(\theta)$ by adjusting the parameters θ of the network. (Goodfellow et al. 2016, Chapter 8). A usual choice in optimization is to have a gradient based method that minimizes the cost function in an iterative manner. A visualization of minimization with gradient based method can be seen in Figure 9.

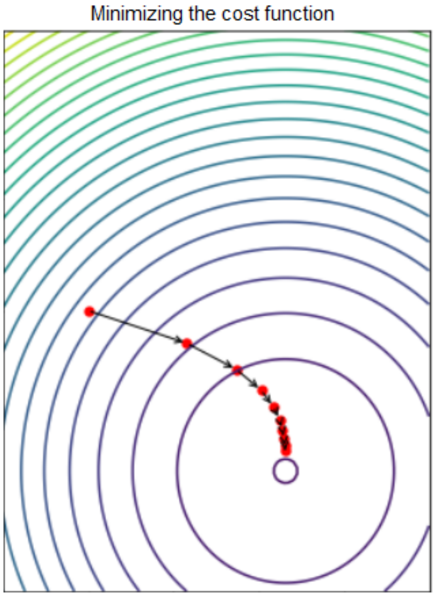

Figure 9. A contour plot of cost function and finding optima by using gradient descent.

The cost function C can be presented in many ways, but a basic form is the following:

$$C(y,\hat{y}) = \frac{1}{m}\sum_{i}^{m} L(y_i,\hat{y}_i), \tag{2.7}$$

where m presents the amount of training samples, L presents the loss function, y presents the desired output, and $\hat{y}$ presents the output predictions of the model. (Goodfellow et al. 2016, Chapter 8). The loss function should be carefully selected for a specific problem, as it has a tremendous affect on the learning of a network. One popular loss function is the cross-entropy loss L_e, where negative logarithm is calculated for the predicted output $\hat{y}$ and for the desired output y by:

$$L_e(y,\hat{y}) = -\sum_{i}^{N} y_i log(\hat{y}_i), \tag{2.8}$$

where N denotes the amount of class labels. (Bishop 2006). The cross entropy loss L_e defines how large the error between the predicted output $\hat{y}$ and desired output y is, and outputs this as a value for each class N between zero and one. It was noted by Simard et al. (2003) that cross-entropy loss can train a model faster and improve the performance of a model when compared to the mean squared error.

Backpropagation is usually described as the most important part of training modern neural networks. The backpropagation algorithm is used to calculate the gradients of the cost function (Nielsen 2015). Rumelhart et al. (1986) was the first who introduced the backpropagation algorithm to be applied in training of neural networks, which enhanced the calculation of gradients drastically. The algorithm for backpropagation can be found from many sources (Rumelhart et al. 1986; Goodfellow et al. 2016), but the algorithm principle, motivated by the work of Nielsen (2015), is performed in the following way. First, the inputs $\mathbf{x}$, the weights $\mathbf{w}$, and the biases $\mathbf{b}$ of the network are initialized. Then, the network is forward propagated through each layer $l = 2,3,...,O$ by computing $\mathbf{s} = \mathbf{w}^l\mathbf{a}^{l-1} + \mathbf{b}^l$, where $\mathbf{a}$ denotes the activations that are calculated by using the activation function f, as seen in the Equation 2.2. When the network has been forward propagated the network outputs the predictions. Next, the predicted outputs and the desired outputs are compared by computing the error δ of the output layer O for each neuron j, by calculating the gradient of the cost function ∇C in the following way: $\delta_j^O = \nabla_{\mathbf{a}} C \odot f'(\mathbf{s}^O)$. The error δ is then calculated for the whole network, starting from the last layer $O-1$ and continuing all the way back to the first layer. The layers from the last layer to the first layer are denoted by $l = O-1, O-2, ..., 2$ and the error of the layers is propagated back by computing: $\delta^l = ((\mathbf{w}^{l+1})^T \delta^{l+1}) \odot f'(\mathbf{s}^l)$, where the weights $\mathbf{w}$ are transposed. The error values show how much the parameters of the network should be adjusted in order to gain more optimized output. Finally, when the error of the network is calculated we can use the chain rule to calculate more optimal values for the networks weights w and the biases b. First, this is calculated with respect to all of the weights $\mathbf{w}$ of the neurons j in the network layers l by computing $\partial C/\partial \mathbf{w}^l_j = \mathbf{a}^{l-1} \delta^l_j$. Then, the rate of change in regard to all of the biases $\mathbf{b}$ of the network layers l is calculated by: $\partial C/\partial \mathbf{b}^l_j = \delta^l_j$. (Nielsen 2015). An optimization method, such as stochastic gradient descent (SGD) or Adam can then use the calculated gradients in order to adjust the parameters of the model. This enables to minimize the difference between the predicted output and the desired output by iteratively adjusting the parameters of the network. (Goodfellow et al. 2016, pp. 200-217).

Stochastic gradient descent (SGD) is a faster and more computationally efficient optimization method when compared to the original gradient descent method. (Bottou 2010; Goodfellow et al. 2016, pp. 149-150). The SGD attempts to iteratively update the cost function into the steepest descending direction trying to reach a local or a global minimum. The update of a step size can be modified with a learning rate parameter of the SGD. (Goodfellow et al. 2016, pp. 290-296). The SGD optimization method for updating the parameters θ of the network is calculated in the following way:

$$\theta = \theta - \eta \nabla_\theta C(\theta; \mathbf{x}, y), \tag{2.9}$$

where η denotes the learning rate, $\nabla_\theta C$ denotes the gradient of the cost function, $\mathbf{x}$ is the training sample, and y is the ground truth. The update of SGD is done by using subsamples from the dataset. (Goodfellow et al. 2016, pp. 149-150). As the learning rate can have a great impact on the learning, more automated methods have been developed. One example is the Adam method (Kingma and Ba 2014), which uses the adaptive learning rate for stochastic optimization. The learning rate in Adam is adjusted by the method itself during the network training. This method can help the optimizer to gain faster convergence as the method is less likely to get stuck in nonoptimal valleys. (Goodfellow et al. 2016; Kingma and Ba 2014).

The optimization problem of deep learning models is quite complex, as the cost function is nonlinear and nonconvex, which means that there are multiple local optima and a global optimum. Minimizing the cost function is conducted by searching a global or a local minimum of the cost function. The global minimum is the point where some function g obtains its absolute lowest value, whereas the local minimum is a point where a function g is lower than any other points nearby (Goodfellow et al. 2016, pp. 80-84). Usually the global minimum can be hard or expensive to solve, thus, in these types of problems the approach is to find the local minimum with a low error rate. The optimization is continued iteratively until the loss changes of the model are very small or the changes stop, meaning that the model has converged (Goodfellow et al. 2016). Furthermore, the global minimum might also lead to overfitting of the model as Choromanska et al. (2014) proved. In addition, Choromanska et al. (2014) also noticed that the local minimum in large networks seem to lie close to the global minimum. Therefore, finding the local minimum seems to produce reasonable results in optimization of deep learning models. (Goodfellow et al. 2016, pp. 279-290).

Regularization of deep learning models

The idea with deep learning is to train a model such that it performs well with unseen data. When a model learns training data well, but is not able to perform well with new data, the model suffers from overfitting. (Goodfellow et al. 2016, Chapter 5). There are few regularization and normalization methods to help with training a model such that it would not suffer from overfitting, such as dropout, regularization terms, data augmentation, and batch normalization.

A dropout layer regularizes a model by randomly dropping out some amount of units (Srivastava et al. 2014). A dropout can help with overfitting but also it is computationally efficient regularization method (Srivastava et al. 2014; Goodfellow et al. 2016, pp. 255-265). Other common types of regularization in deep learning models are the L1 and the L2 regularisation that add penalties to the weights of the network (Goodfellow et al. 2016, pp. 225-233).

As mentioned before, deep learning models require a large amount of data in order to generalize. A common problem especially in biomedical imaging is that the amount of data is limited. Data augmentation is a method to synthetically create more training data (Simard et al. 2003; Goodfellow et al. 2016, p. 236-238). Specifically data augmentations can increase the models capability to learn with more variance and to reduce overfitting (Simard et al. 2003; Goodfellow et al. 2016, pp. 236-238). Although, real data have been proven to perform better than augmented data, augmentations can still help the model to learn better (Wong et al. 2016; Xu et al. 2016). Data augmentation also have undesired effects when used carelessly, bad augmentations may actually decrease the performance of a model (Wong et al. 2016). Therefore, data augmentations can be useful to create variance into model learning, but the results with different models may vary.

Another method to help with training deeper models to generalize better is a method called batch normalization (Goodfellow et al. 2016, pp. 313-317). As the parameters of the network change iteratively during training, these changes can have a huge impact on the learning of a model. To avoid the impact of imbalanced parameters, the batch normalization can be applied to the layers of a model. The basic principle of a batch normalization layer is to normalize the activations of a model in a specified layer. (Ioffe and Szegedy 2015). Batch normalization was first introduced by Ioffe and Szegedy (2015), and it consists of the following steps. First, the mean of a batch $\bar{B}$ is calculated with the batch size c and the activations $\mathbf{a}$ by: $\bar{B} = \frac{1}{c}\sum_{i=1}^{c}\mathbf{a}_i$. Next, the variance V of a batch B is computed with respect to the activations a and the mean of batch $\bar{B}$ by $V_B^2 = \frac{1}{c}\sum_{i=1}^{c}(\mathbf{a}_i - \bar{B})^2$. Then, the activations $\mathbf{a}$ are normalized as $\hat{a}_i$ by using: $\hat{a}_i = \frac{\mathbf{a}_i - \bar{B}}{\sqrt{V_B^2 + \varepsilon}}$, where ε denotes a constant. Finally, the normalized data $\hat{a}_i$ is shifted and scaled by using learnable parameters γ and β, and the output of batch normalization h_i is conducted by: $h_i = \gamma\hat{a}_i + \beta$. (Ioffe and Szegedy 2015).

Cross-validation of deep learning models

As we have presented the training and the regularizing methods of a deep learning model, we will now continue on discussing how to validate the results of a model. In order to verify the models prediction capabilities the dataset is divided into separate sets. Dataset needs to be divided into training and test set, so that the model can be tested with unseen data that have not been used in the training phase of the model. For hyperparameter tuning a validation test set is also needed. (Goodfellow et al. 2016, pp. 118-120). A usual procedure is to split the data into training, validation, and test sets. With a small dataset this can be problematic, as the model has fewer examples to learn from. To address this issue, cross-validation can be applied. Cross-validation divides the data into subsets that are used to train and test the model, from which the network performance can be evaluated. (Goodfellow et al. 2016, pp. 118-120). Another problem when working with unbalanced dataset is to split the classes of the data evenly into separate subsets, such that all of the subsets would contain elements from all of the classes. The stratified cross-validation is able to handle this problem of dividing unbalanced dataset. It splits the dataset in a way that all of the folds contain all class labels quite evenly.

Most common cross-validation method is the k-fold cross-validation (Goodfellow et al. 2016, pp. 118-120). In the k-fold cross-validation the data is split into k amount of subsets, where $k-1$ of folds are selected as the training set and the remaining fold is selected as the test set. The model is repeatedly trained k amount of times with the different folds. Finally, the performance of the model is reported as the average result from all of the k amount of trials. (Goodfellow et al. 2016, pp. 118-120).

2.3.3 Convolutional neural networks

Having defined the basics of deep learning and ANN's, we will now move on to discussing a specialization of deep neural networks, called convolutional neural networks. CNN is a modern type of deep learning model that enables an efficient way of training models with a large amount of data, and the models can even handle unprocessed data quite well. (Krizhevsky et al. 2012).

The idea of convolutional neural networks evolves from a study of the visual system of the brain by Hubel and Wiesel (1962), where they measured the response of neurons with visual stimuli from a cat's brain. They found that the neurons of receptive fields had specialized layers that were able to detect different features. They also found that these layers have hierarchy - lower level layers detected simple features, whereas higher levels layers had more complex feature detection. (Hubel and Wiesel 1962). The idea of building an artificial neural network with similar hierarchical layers was first introduced by Fukushima (1980). He developed the neocognitron model, and this network architecture idea was the core of convolutional neural networks (Fukushima 1980). Later this model was modified by LeCun et al. (1989) when they introduced a CNN trained with a gradient-based learning algorithm, backpropagation, to recognize handwritten digits. But it was only in 2012 when CNNs achieved their breakthrough in deep learning, when Krizhevsky et al. (2012) introduced their version of CNN, the AlexNet. This model had outstanding results in image classification on ImageNet competition in 2012 (Russakovsky et al. 2015). Their network was deeper and larger than the earlier models and it was able to make use of large amount of data. The training of the model was fast, as the model was trained by using GPUs. (Krizhevsky et al. 2012). They also generated new training data examples by using data augmentations and used the dropout regularization method. This network architecture helped to reduce the model from overfitting. (Krizhevsky et al. 2012).

After the work of Krizhevsky et al. (2012) CNNs became widely used in deep learning. (Goodfellow et al. 2016, pp. 365-366). A study by Simonyan and Zisserman (2014) evaluated deeper CNN models with a mission to improve the accuracy of the CNN models. They came to the conclusion that the deeper the CNN model is the better the accuracy. By increasing the depth of the CNN model they were able to gain state-of-the-art results in classification problems and localization problems. (Simonyan and Zisserman 2014). Next, the main features of CNNs are introduced - the convolutions and the pooling operations.

Convolution

The usage of convolutions allow the models to learn separate spatial information from the data. Convolutions create feature maps that represent different features from the training data. The convolution operation used in CNNs is the convolution operation that does not use the kernel flip, and its two dimensional form can be calculated by:

$$\mathbf{S}(o,p) = (\mathbf{K} * \mathbf{X})(o,p) = \sum_{r}\sum_{v}\mathbf{X}(o+r,p+v)\mathbf{K}(r,v), \tag{2.10}$$

where $\mathbf{S}$ denotes an output feature map with dimensions of o and p, $\mathbf{X}$ denotes a two-dimensional input data, and $\mathbf{K}$ denotes a two-dimensional kernel of weights with dimensions of r and v. (Goodfellow et al. 2016, pp. 327-329). The basic convolution operation is visualized in Figure 10. In convolution we have a kernel $\mathbf{K}$ that is slid across the input, as seen in the figure. After the kernel has been slid in each location of an input data **X**, we get a feature map **S** as an output of the convolution operation. Each feature map is different, as the weights of a kernel are unique. Therefore, each feature map is able to learn different features from the data. (Dumoulin and Visin 2016). The amount of elements the kernel is slid can be modified. Also, the boundaries of the input array can be padded, otherwise the feature maps are downsampled with each convolution operation when compared to the input data. (Goodfellow et al. 2016).

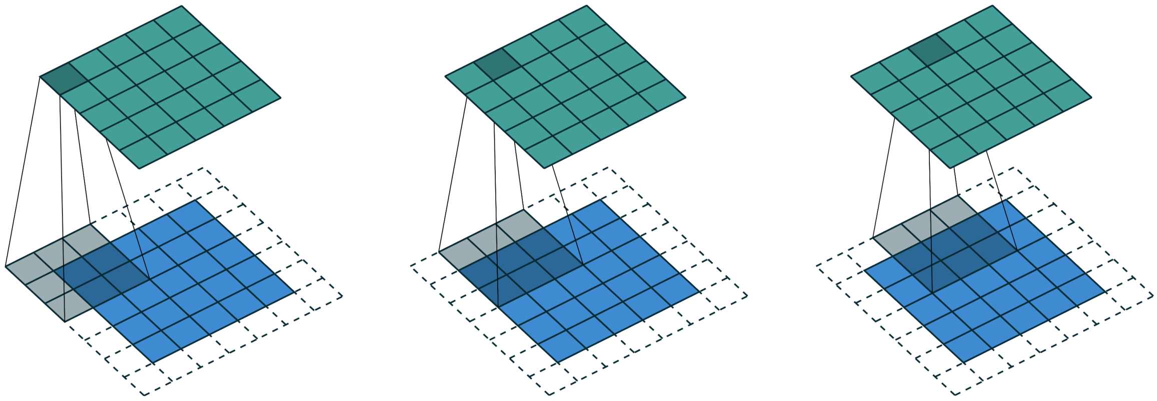

Figure 10. Convolution operation with padding (reproduced from Dumoulin and Visin (2016)). The input feature map is presented as blue, and the kernel is presented as the grey shadow. The output feature map is presented by the green color. The convolution operation is applied with a stride of 1 and a kernel size of 3×3.

Up-convolution

Up-convolution is also known as the transposed convolution. In up-convolution the feature maps are upsampled with learnable parameters (Dumoulin and Visin 2016; Long et al. 2015). This operation allows the network to upscale the feature maps, for instance to gain the same dimensions as the input data (Dumoulin and Visin 2016). This method was noted to be efficient in segmentation tasks, as the upsampling is based on learning from the training process of a network (Long et al. 2015). In up-convolution the forward and the backward passes are the opposite when compared to the normal convolutions, as the operation is performed by transposing the convolutions. (Dumoulin and Visin 2016). A visualisation of the up-convolution method can be seen in Figure 11.

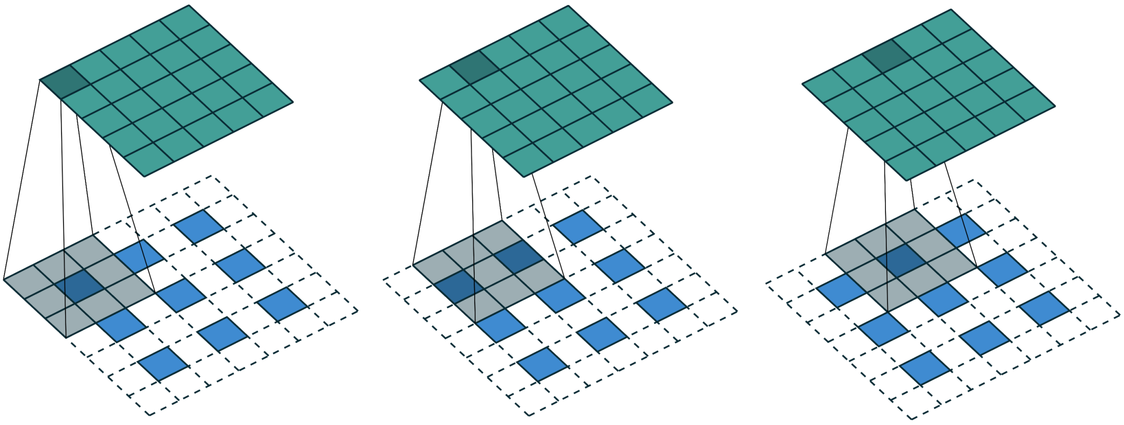

Figure 11. Up-convolution operation with a 3×3 kernel, a padding of 1, and a stride of 2 (reproduced from Dumoulin and Visin (2016)). The input feature map is presented as blue, and the kernel is presented as a grey shadow. The output feature map presents the increased spatial dimensions, that is presented by the green color.

Pooling

In CNNs it is typical that convolution is followed by a pooling layer. The pooling layer reduces the dimensions of the feature map, which in addition reduces the amount of computation. (Dumoulin and Visin 2016). The pooling also makes the model more invariant to spatial translations of the input data. (Goodfellow et al. 2016, pp. 335-339). The pooling is applied by sliding a kernel across the feature map. The step size, also known as the stride, of a pooling operation can be modified. The output of the pooling is a downsampled feature map. (Dumoulin and Visin 2016).

There are several methods to compute the pooling, such as the max pooling and the average pooling, which are visualized in Figure 12. In max pooling the kernel is slid across the input feature map, and within each step a maximum value is selected inside from the neighborhood of the kernel. The average-pooling, on the other hand, calculates the average of the items in the neighborhood within each step of the kernel sliding across the feature map. (Goodfellow et al. 2016, pp. 335-339).

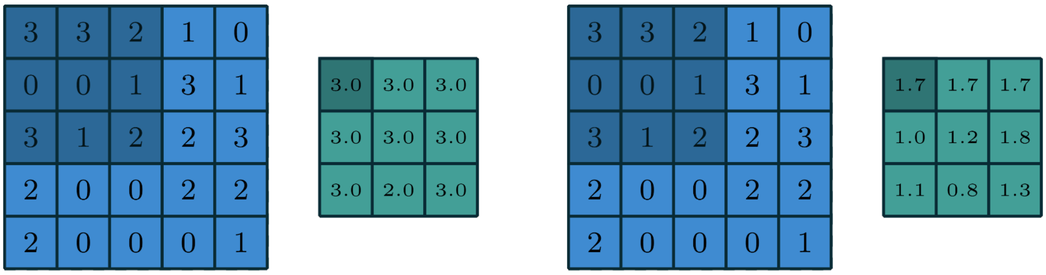

Figure 12. Pooling methods visualized (reproduced from Dumoulin and Visin (2016). On the left side the max pooling operation is visualised and on the right side the average pooling operation is presented. Both pooling methods have a kernel of size 3×3 and a stride of 1. The blue presents the input feature map, the dark blue is the kernel to be slid accross the input feature map, and the green presents the output of the pooling operation.

2.4 Semantic segmentation

Having introduced the basics of deep learning, we will now continue on discussing the predictive methods of deep learning models, having the main focus on semantic segmentation. We will introduce the state-of-the-art deep learning models, together with different architectures and building blocks. Finally, we will focus on semantic segmentation in biomedical problems, especially focusing on skin cancer segmentation.

Deep learning models that are trained with images have several methods to output predictions, such as by image classification, by object detection, or by semantic segmentation. These methods are visualized in Figure 13. Image classification refers to the model outputting a predefined class for the whole image. The model can also output multiple classes for an image, which is referred as multi-label classifier. Object detection models, on the other hand, produce output labels by locating the classes from the input images. Whereas, in semantic segmentation a neural network outputs a pixel level segmentation map from an input image. Each pixel of the output segmentation map is labeled into predefined categories. In this study we will focus on semantic segmentation in deep learning.

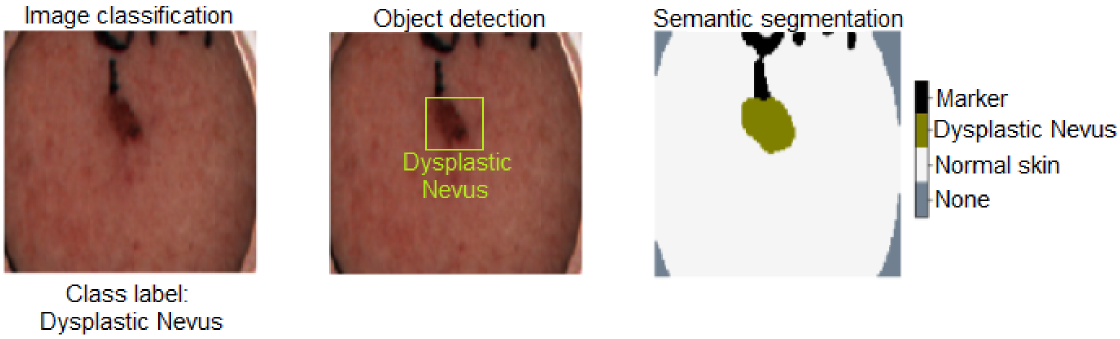

Figure 13. Example outputs of image classification, object detection, and semantic segmentation.

Semantic segmentation models are trained by using a training image and a ground truth image. Ground truth represents the correct classes for each pixel in the original image. The output of the semantic segmentation model is a pixel-wise classification. An example of the semantic segmentation model input image, ground truth labels, and an output image are seen in Figure 14.

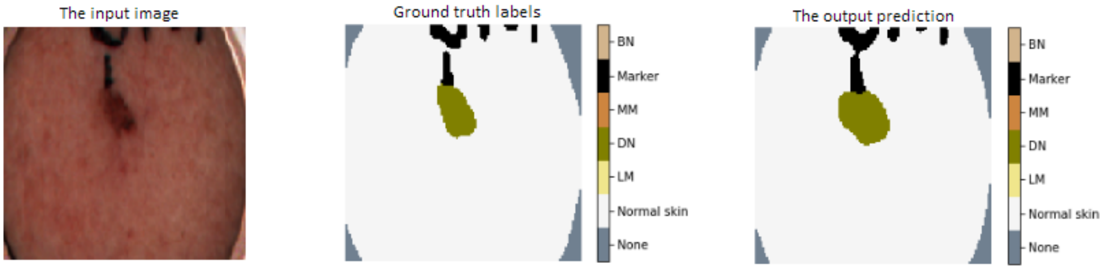

Figure 14. An example of semantic segmentation data: an input image, ground truth data, and an output prediction. The input image is segmented by the model in pixel-wise to different categories, such as normal skin, marker, and lesion types. The output prediction map of a segmentation model produces classification labels for each pixel.

Semantic segmentation has been applied in several fields, such as in scene understanding (Badrinarayanan et al. 2015), in remote sensing (Henry et al. 2018), and in segmentation of biomedical images (Ronneberger et al. 2015). Semantic segmentation is a highly active research field and the development of models is rapid (Alom et al. 2018). For example, Long et al. (2015) developed fully convolutional networks (FCNs) for segmentation, where all the layers of the CNN are convolutional layers. With this architecture they were able to improve the results of the previous state-of-the-art models. (Long et al. 2015). Another framework in semantic segmentation that brought superior results was the Segnet model by Badrinarayanan et al. (2015). They presented an encoder-decoder architecture in segmentation and were not only able to improve the road scene understanding results but also they improved the speed and memory usage of the model (Badrinarayanan et al. 2015).

An important feature in the encoder-decoder architecture is its capability to maintain localization information from input images into output segmentation (Ronneberger et al. 2015; Badrinarayanan et al. 2015). The encoding phase of the architecture downsamples the dimensions of the data. This enables faster computing time and decreases the memory usage. Whereas, the decoding phase upsamples the image data dimensions. (Badrinarayanan et al. 2015). Upsampling of the data can be performed, for instance by using up-convolutions (Ronneberger et al. 2015). The architecture of encoder-decoder can be modified to use any CNN model, for example AlexNet (Krizhevsky et al. 2012), VGGNet (Simonyan and Zisserman 2014), or ResNet (He et al. 2015a). (Siam et al. 2018; Badrinarayanan et al. 2015).

In biomedical segmentation an encoder-decoder architecture, called the Unet by Ronneberger et al. (2015), became popular after it won the EM segmentation challenge (Arganda-Carreras et al. 2015) in 2015. In biomedical problems it is usual to have a small dataset. The Unet model was built to tackle this problem and succeeded to gain good results even with a small dataset. The Unet model uses convolutional blocks in both downsampling and upsampling the feature maps. The feature map dimensions are increased after each convolution block. The architecture uses skip connections to pass cropped feature maps after each convolution block from encoding side to decoding side. The Unet has been used in semantic segmentation with a wide range of tasks, such as in biomedical image problems (Gorriz et al. 2017; Alom et al. 2018), in remote sensing (Zhang et al. 2018) as well as in scene understanding (Siam et al. 2018). The encoder-decoder network, combined with skip connections between layers, is able to maintain details between the layers that otherwise are lost. (Ronneberger et al. 2015).

The deeper the models get the harder the training becomes. To address this problem, He et al. (2015a) introduced the ResNet architecture. They used residual connections, which are a type of skip connections between every few weight layers to pass trough information between the layers. The mathematical form of the residual connection G is the following:

$$G(\mathbf{x}) = W(\mathbf{x}) + \mathbf{x}, \tag{2.11}$$

where $W(\mathbf{x})$ presents the nonlinear mapping learned in a layer, and $\mathbf{x}$ denotes the input features (He et al. 2015a). The ResNet contain several weight layers that consist of convolution operations, batch normalization, and ReLu operation. The feature map dimensions of weight layers are increased after a few weight layers. The ResNet architecture won several competitions in classification, in object detection, and in segmentation. It was shown that residual functions boost the models optimization but also they improve the accuracy of a model. (He et al. 2015a). The ResNet model and especially its residual connections have been successfully adopted in segmentation tasks, for instance by combining them with building blocks of the Unet architecture (Drozdzal et al. 2016; Milletari et al. 2016; Siam et al. 2018; Zhang et al. 2018).

2.4.1 Semantic segmentation in skin cancer detection from hyperspectral images

In this thesis we focus on a specific medical imaging problem – the semantic segmentation of four different lesion types BN, DN, LM, and MM. The detection of skin cancer can be problematic, therefore a novel tool is needed to help in clinical examinations, as discussed in Section 2.1.2. Several deep learning methods have been used to detect skin cancer but most of the studies were conducted by using RGB images, multispectral images, or their combinations (Alom et al. 2018; Esteva et al. 2017; Gorriz et al. 2017; Tomatis et al. 2005; Yu et al. 2017). There are less studies from skin cancer detection by using hyperspectral images (Johansen et al. 2020). A novel hyperspectral imaging system is potential tool to help in skin cancer examinations, but further development is needed, for example to automate the imaging results (Neittaanmäki-Perttu et al. 2013; Neittaanmäki-Perttu et al. 2015; Salmivuori et al. 2019).

Hosking et al. (2019) used a hyperspectral dermoscopy that was able to image lesions with

21 different wavelengths, from the range of 350 – 950 nm. The HSI images were analyzed by using Monte Carlo simulation method, with successful results of 100 % recall (also known as sensitivity), and of 36 % specificity in melanoma detection (Hosking et al. 2019). Furthermore, Gu et al. (2018) released a hyperspectral dermoscopy dataset of 330 lesions. The images of the dataset have dimensions of 256 × 512 pixels and contain 16 spectral bands from the range of 465 – 630 nm. The study of Gu et al. (2018) used support vector machine (SVM) to classify different lesions from melanoma, and gained 84 % recall and 72.10 % specificity in melanoma detection. Another study by Fabelo et al. (2019) used a dataset of 49 hyperspectral images with a size of 50 × 50 × 125 pixels. The imaged wavelength regions were from the range of 450 – 950 nm. They used SVM classifier and obtained results with a mean of recall of 75.33 % including all lesion types, but only one melanoma lesion was correctly classified. (Fabelo et al. 2019). These recent studies suggest that hyperspectral imaging combined with automated classification is a promising method in skin cancer detection, but further investigation is needed. Compared to the previous studies, this study used a novel HSI dataset that contained 120 different contiguous wavelengths with image dimensions of 1920 × 1200 pixels with 61 images of lesions.

There are not many studies that have used hyperspectral imaging combined with deep learning and semantic segmentation of lesions. A recent study by Pölönen et al. (2019) collected novel HSI data and compared different architecture structures of CNNs – using 1D, 2D, and 3D convolutions. The results showed that the performance of a CNN can be increased with the usage of spectral and spatial data. The results were promising and varied between different models. Recall varied from 93 % to 100 %, whereas specificity varied from 12 % to 21 %. Finally, precision metric that is also known as positive predictive value, varied from 32 % to 34 %. Pölönen et al. (2019) conclude that the results of some models were able to outperform clinicians, whereas some models maintained the same accuracy in diagnostics when compared to the study by Heal et al. (2008). However, the models used in the study had some difficulties with segmentation accuracy, especially in the border areas. (Pölönen et al. 2019).

In this study we continue to work with the novel hyperspectral image data in order to improve the semantic segmentation of lesions and lesion classification. The aim is to compare two deep semantic segmentation models: a slightly modified original Unet model (Ronneberger et al. 2015), and a modified ResNet model (He et al. 2015a) combined with building blocks of the Unet model (ResNet-Unet). As the encoder-decoder architecture has produced good results in previous biomedical semantic segmentation tasks with RGB and multispectral data (Alom et al. 2018; Drozdzal et al. 2016; Zhang et al. 2018), we are encouraged to test the performance of this architecture with the novel hyperspectral data.

3 Materials and methods

In the previous chapter the background and the main concepts of this research were briefly introduced. In this chapter we present the materials and methods used in the study. First, in Section 3.1 the materials of the study are introduced. Then, the methods and experimental setup is presented in Section 3.2. Finally, the evaluation of the results are viewed in Section 3.3.

3.1 Materials

Next, we will present the data that was used in this study. Then, we describe the preprocessing techniques we used for the data. Finally, we see the annotation process for the training data, which create the ground truth class labels for each pixel in the lesion image data.

This study used a novel dataset of hyperspectral images. The dataset used was small and contained only 61 images of lesions. The dataset contained four different lesion types – benign nevus, dysplastic nevus, lentigo maligna, and malignant melanoma. The dataset did not contain an even amount of different lesion types. The amount of each lesion type examples the dataset contained are presented in Table 1. An example from each lesion type can be seen in Figure 15. The dataset also contained a histopathologically verified result, one lesion type diagnosis for each image of a lesion. This gave a verified lesion type for each image, that was used as the true label when comparing the prediction results of the implemented models.

Lesion type	Amount of examples in the dataset
Benign Nevus	14
Dysplastic Nevus	26
Lentigo Maligna	6
Malignant Melanoma	15

Table 1. The amount of lesion type examples the used dataset contained.

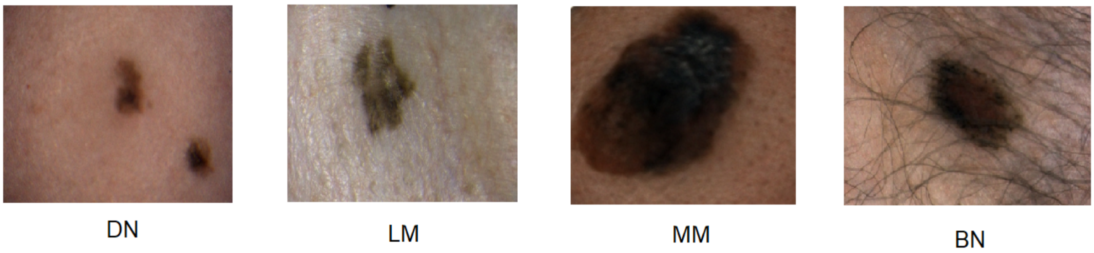

Figure 15. Examples of lesion types used in the study: Dysplastic nevi (DN), Lentigo Maligna (LM), Malignant melanoma (MM), and Benign nevi (BN). The images are from the dataset that was used in this study.

The novel HSI data used in this study had been collected by Pölönen et al. (2019) in 2016–2017 from volunteer patients. The data were gathered from two hospitals: from the Department of Dermatology and Allergology of Helsinki University Hospital, Helsinki, Finland, and from the Päijät-Häme Central Hospital, Lahti, Finland. This study had selected only patients whose lesions were to be excised and examined histopathologically. The more detailed description of the data collection can be found from a study by Pölönen et al. (2019).

The study by Pölönen et al. (2019) gathered the novel dataset of hyperspectral images of lesions, by using two novel and identical hyperspectral imaging system prototypes (Prototype 2016 by Revenio Group, Finland). These novel HSI prototypes used Fabry-Pérot interferometer (FPI) to separate the wavebands. In addition, external illumination was blocked by using a covering tube on the imagers, which was seen in some of the image borders in the novel dataset. To create diffuse illumination to imaging of the lesions, the novel HSI imagers had been integrated with a diffuse lighting. The specifications of the novel hyperspectral imaging system prototypes are available in Table 2. (Pölönen et al. 2019). All in all, the total image dimensions that the prototypes were able to capture were $1920 \times 1200 \times 120$ pixels. The more detailed description of the novel HSI prototypes and the data can be found from the study by Pölönen et al. (2019).

Having introduced the novel dataset used in the study, we will now continue to observe the preprocessing of the data. The image data were in the format of radiance. In order to use the spectral information and inspect the spectral signatures of the objects the data contained,

Imaging specifications	Details
Amount of wavebands in images	120
Full width at half maximum (FWHM)	5–15 nm
Sensor	CMOS
Spatial resolution of the images	15 μm/pixel
Spectral separator	Fabry-Perot interferometer (FPI)
Spectral range of wavebands	450–850 nm
Resolution of the machine vision camera	1920 × 1200 pixels

Table 2. Hyperspectral imaging systems specifications used for the image capturing. The more detailed description of the imaging system can be found from Pölönen et al. (2019).

the data were converted from radiance to reflectance by using Equation 2.1 introduced in Section 2.2.2. For this operation, the dataset also contained a white reference data cube for each image.

Unfortunately, the data contained some imaging artefacts in the last 20 wavelengths, therefore these wavelengths were removed from the image data (Pölönen et al. 2019). After the operation, the image data contained 100 different wavelengths. The study by Johansen et al. (2020) suggests that dimensionality reduction with HSI data can improve the performance, remove data redundancy, and remove noise the data contains. Therefore, we removed every other wavelength from the 100 wavelengths, as the data contained spectral overlap. All in all, the dimensions of the downsampled hyperspectral data cube were $128 \times 128 \times 50$ pixels.

As we have now discussed about the data and the preprocessing of the data, we will move on to describe the annotation process. Annotation was a significant aspect in training the experimental models, as they present the ground truth labels for the images of lesions. The hyperspectral data cubes were annotated by using the MATLAB R2018a Image Labeler toolbox. We used full images in annotation and in training of the deep learning models. According to Johansen et al. (2020) the usage of full images that utilize also spatial information, is a realistic method as it is the method of how clinicians view the lesions. In addition, this may bring more information of the surroundings and improve the performance of the models (Johansen et al. 2020). The images were labeled with seven class labels: malignant melanoma, lentigo maligna, dysplastic nevus, bening nevus, none, marker, and normal skin. Each data cube contained an image of a lesion, which was annotated to a one lesion type. The ground truth diagnosis for the lesion an image contained, was obtained from the histopathologically verified result. Some of the images contained borders from the imagers covering tube, as mentioned earlier. These pixels we annotated to the label none. In addition, some of the images contained marker drawings in the skin, and these pixels were labeled to the marker class. Finally, the skin area pixels were annotated as the normal skin class, and to this class we also included the hair on the skin. An example of ground truth labeling can be seen in Figure 16. The annotations were labeled by a non-specialist.

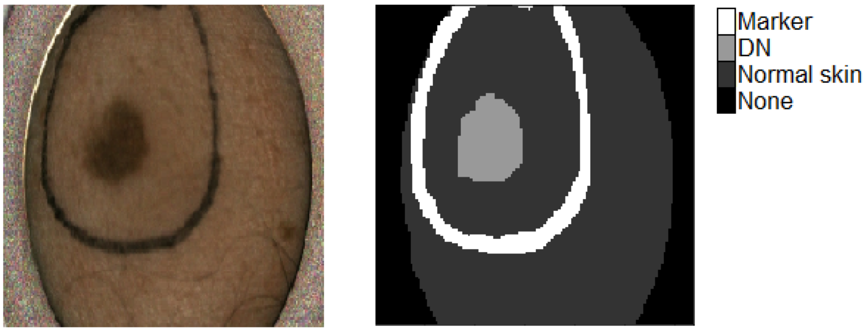

Figure 16. A false color hyperspectral image and an annotated ground truth image of a lesion that has been histopathologically verified as DN. Other classes that the region of interest (ROI) contained are marker, normal skin, and none for the covering tube visible in the corners of the image.

3.2 Methods

Having discussed about the materials of the study, we will now move on to introduce the methods used in the study. The experiments were performed by implementing two different deep learning models. The models were trained with the hyperspectral image data together with the annotated image data, which were introduced in Section 3.1. In this section we will first present the architectures of the implemented semantic segmentation models, and finally we will report the experimental setup used in the study.

The two semantic segmentation architectures that we implemented in this study made use of the encoder-decoder structure. In addition, the other architecture also used the residual connections. This was due to that many studies have used the encoder-decoder structure with promising results in semantic segmentation tasks (Ciresan et al. 2012; Badrinarayanan et al. 2015; Ronneberger et al. 2015). A major advantage of using the Unet model, is its ability to generalize well even with small datasets (Ronneberger et al. 2015). Also, the residual connections in the ResNet model (He et al. 2015a) have allowed to build even deeper models with successful results (He et al. 2015a). Recent studies have also combined the Unet model and the ResNet model with good results (Zhang et al. 2018; Siam et al. 2018; Alom et al. 2018). Therefore, in this study we implemented a slightly modified Unet model as the first model. As the second model, we implemented a modified ResNet model combined with the Unet model (ResNet-Unet). Next, we will introduce the two implemented semantic segmentation models.

1. **Unet**
   The first semantic segmentation model we implemented in this study was a modification of the Unet model (Ronneberger et al. 2015). The original Unet model (Ronneberger et al. 2015) was modified to work with hyperspectral data, such that the input layer accepted the input data cubes in the size of $128 \times 128 \times 50$ pixels. In addition, we used padded convolutions, in order for the feature maps to preserve the same dimensions as the input features. As a result, cropping was removed from the architecture. Also, the batch normalisation layers were added to the network on the input layer and after each convolution block. Finally, the output predictions of the model were modified to output multi-class segmentations. All in all, the implemented Unet model contained only a few modifications and the basic structure was kept in similar form as it was presented in the original paper by Ronneberger et al. (2015). Figure 17 presents the architecture of the implemented Unet model.

2. **ResNet and Unet model combination (ResNet-Unet)**
   The second architecture used in this study was a modified ResNet model (He et

al. 2015a) combined with the Unet model (Ronneberger et al. 2015), which we refer to as the ResNet-Unet model. First, the input layer of the original ResNet-34 model (He et al. 2015a) was modified to be able to have the input data in the size of $128 \times 128 \times 50$ pixels. Next, we modified the original ResNet-34 model to output semantic segmentation masks. We also combined the ResNet model with the Unet (Ronneberger et al. 2015) model. We adapted from the Unet model its u-shaped encoder-decoder structure and also its skip connections between the encoding and decoding phases. We implemented the skip connections after each of the four different dimensional convolution blocks of the ResNet. A presentation of the implemented model can be seen in Figure 18. In addition, the convolution blocks were modified such that all of the four blocks contained six convolutional layers. After each 3×3 convolution operation a batch normalization and ReLU activation was applied. A presentation of the convolutional layers are seen in the bottom of Figure 18. Also, the fully convolutional layer and the pooling layer were removed from the end of the ResNet-34 model. Otherwise, on the encoding phase the ResNet-34 model was not modified compared to the original model. After the encoding phase, the decoding phase was implemented by using the ResNet-34 model similarly as in the encoding phase. On the contrary, the implemented decoding ResNet-34 network upsampled the features by using a 3×3 up-convolution. The final layer of the model had a 2×2 up-convolution with a stride of 1. This was followed by a 3×3 padded convolution layer. Finally, the features were passed to the softmax activation function to produce the output segmentation maps.

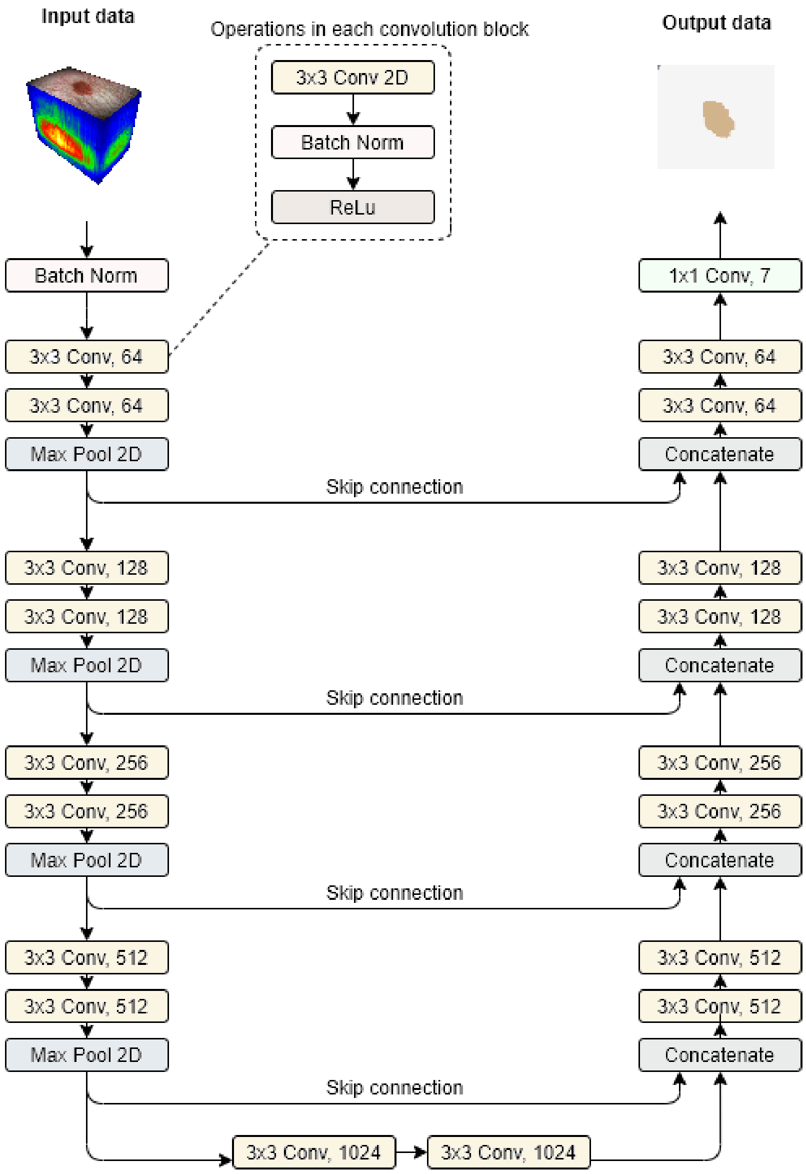

Figure 17. The architecture of the implemented Unet model. The left side presents the encoding phase and the right side shows the decoding phase of the model. The model takes as an input the image data cube and outputs the segmentation map prediction. The encoding and decoding phases were connected with skip connections. In addition, the figure presents a detailed description of each convolution operation.

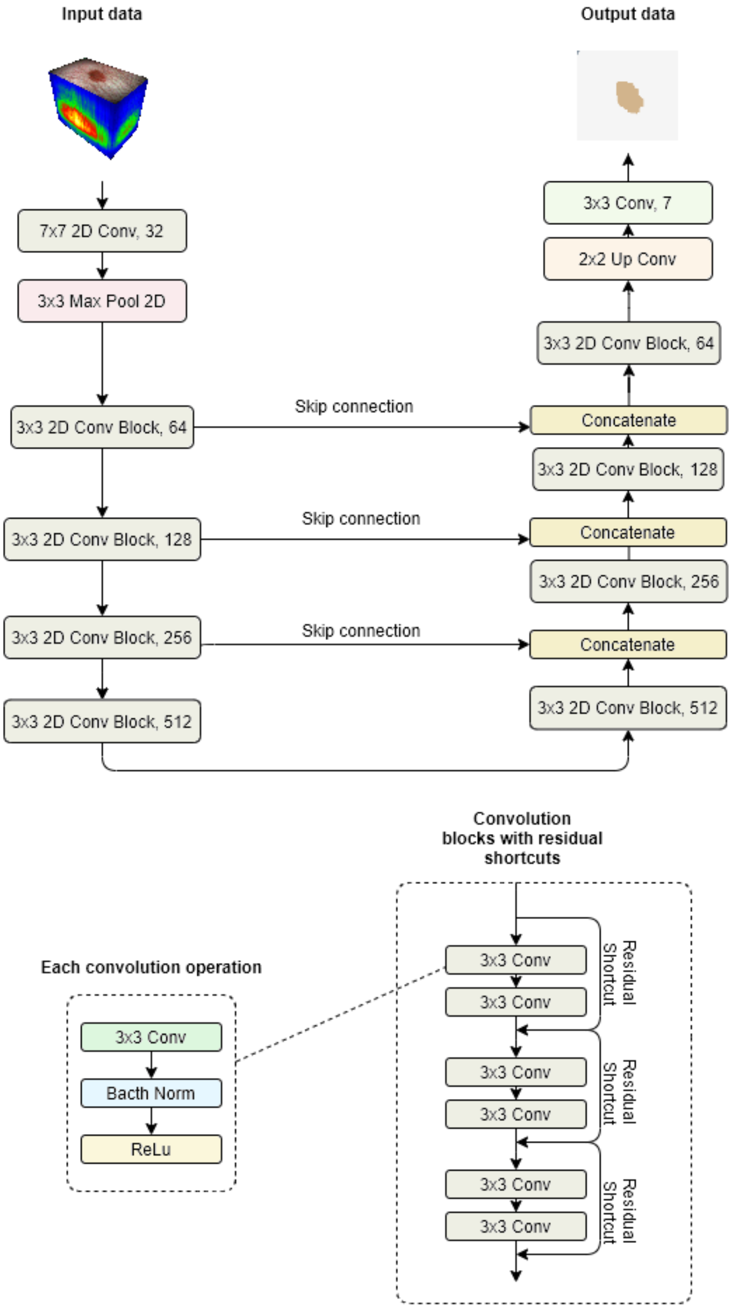

Figure 18. The architecture of the implemented ResNet and Unet model combination (ResNet-Unet). The left side of the graph presents the encoding phase, and the right side of the figure presents the decoding phase of the model. The encoding phase and the decoding phase are connected with skip connections. At the bottom of the figure, the convolution blocks and their residual connections are presented in more detail. Each convolution was followed by the batch normalization and the ReLU activation.

3.2.1 Experiments

As we have introduced the architectures of the models we implemented in this study, we will move on to describe how the experiments were performed. First, we will introduce the environment of the experiments, and then we continue to discuss the workflow of the study.

The environment that the two implemented models used in training were the NVIDIA GeForce GTX 960 with 16 GB RAM and CUDA Version 10.0.130. The hyperspectral data were preprocessed by using Spectral Python (Boggs 2014). The models were implemented by using Python 3.6.5 (Van Rossum and Drake 2009) programming language and by using Keras 2.2.5 (Chollet et al. 2015) deep learning framework with Tensorflow 1.14 (Martín Abadi et al. 2015) back-end.

As we discussed earlier, this study implemented two deep learning models for semantic segmentation, in order to segmentate and classify lesions from the novel hyperspectral image data. The data were preprocessed and the data were annotated with the methods described in Section 3.1. The size of the data after the preprocessing were $128 \times 128 \times 50$ pixels. The predictive performance of the models in this study were estimated by performing a 5-fold stratified cross-validation. This method was necessary to obtain all lesion types to appear in all of the folds, as the dataset contained an unbalanced amount of different lesion types (see Table 1).

The novel HSI dataset was small but deep learning models need a lot of data in order to perform well. Therefore, data augmentations were used to create synthetic data from the original data. The augmentation methods used in the study were: rotation, horizontal flip, vertical flip, cropping, and shifting. The augmented synthetic image data were created by randomly selecting two out of the five augmentation methods described above. Furthermore, the amount of applying variation for an augmentation method was selected randomly. The variation for rotation was applied randomly from the range of $[0.1, 179.0]$ degrees. The images were randomly flipped either horizontally or vertically. The amount of pixels cropped from an image varied from the range of $[0.0, 40.0]$ pixels. The shift of an image was selected randomly from the range of $[0.0, 20.0]$ pixels. Both the image data cube and the annotated mask were augmented together with the same methods and the same variation amounts.

The data augmentations were performed from the original data during training, and individually for each fold in the 5-fold stratified cross-validation process. In the 5-fold stratified cross-validation process the original data had been divided into training set, validation set, and test set with a quite balanced presentation of each lesion type classes in each fold. After this division, we applied the data augmentation methods to create synthetic data to expand the training data set. We only expanded the training data by using data augmentation, but the validation set and the test set contained only the original data. Therefore, the final amount of data within each fold were approximately 50 000 images in the training set, 4 images in the validation set, and 12 images in the test set.

Finally, from Table 3 we can see the amount of parameters the implemented models contained. The parameters of the implemented models were initialized by using He normal (He et al. 2015b) initialization. Furthermore, the models were trained by using the Adam optimizer with a learning rate of 0.001. The models used the categorical cross-entropy as the loss function. In addition, the softmax activation function was used in the final layer with both of the models. All in all, the networks were trained by using 1000 epochs with a batch size of 10.

Model	Trainable parameters	Non-trainable parameters	Total parameters
Unet	31,109,483	12,004	31,121,487
ResNet-Unet	10,984,999	28,672	11,013,671

Table 3. The summary of the amount of parameters the models contained.

3.3 Evaluation

Previously in this chapter we introduced the used dataset, presented the implemented models, and reported the experimental setup of this study. The following part of this chapter moves on to describe the evaluation methods of the models. We will first go through the evaluation of the segmentation and the used evaluation metrics. Finally, we will focus on the evaluation of the lesion classification.

3.3.1 Evaluation of semantic segmentation

The first method of evaluating the implemented models, was the evaluation of the semantic segmentation results. We included all of the seven class labels (BN, DN, LM, marker, MM, none, and normal skin) into evaluating the segmentation results. All in all, we evaluated the overall quality of the segmentation within all of the classes by using the following metrics: precision, recall, f1-score, and specificity. In addition, the segmentation predictions were evaluated visually with respect to the ground truth annotations, to inspect the overall segmentation results. The metrics of precision, recall, and specificity were used in the evaluation of the models, as these metrics are regularly used in biomedical diagnosis tasks (Parikh et al. 2008).

First, precision metric, also known as the positive predictive value (PPV), was calculated in the following way:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}. \tag{3.1}$$

As we can see from the equation, precision evaluates the correct positive cases with respect to all of the predicted positive cases.

Next, we used recall metric, also known as sensitivity in the following way:

$$\text{Recall} = \frac{\text{true positives}}{\text{false negatives} + \text{true positives}}. \tag{3.2}$$

Recall measures the difference between all of the correct positive cases with respect to the all actual positive cases, including the positive cases which have been predicted falsely as negative.

F1-score evaluation metric was calculated in the following way:

$$\text{F1-score} = 2 * \frac{\text{precision * recall}}{\text{precision + recall}}. \tag{3.3}$$

As we can see, the f1-score takes the harmonic mean from precision and recall. It presents an equally weighted comparison of these metrics. The f1-score, or more commonly the dice coefficient metric, is one of the many methods to evaluate the overall segmentation accuracy, and it has been widely used especially in medical volumetric segmentation. (Taha and Hanbury 2015).

Finally, specificity metric, which is also known as the true negative rate, was calculated by:

$$\text{Specificity} = \frac{\text{true negatives}}{\text{false positives} + \text{true negatives}}. \tag{3.4}$$

Specificity evaluates the correctly predicted negatives cases with respect to all actual negative cases.

3.3.2 Evaluation of lesion classification

The second evaluation method for the models was the evaluation of lesion classification, which predicted one lesion type class for each image. The whole image was classified regarding to the most severe tumour or lesion type that was found from the prediction. Table 4 presents the risk classes for each lesion type, from the most severe skin cancer type to benign lesion type. Therefore, if a prediction of a lesion contained all lesion types: BN, DN, LM, and MM, the image was classified as MM, as it has the highest risk class as seen from Table 4. This was due to the fact that malignant melanoma has a high mortality rate, and detecting even a small amount of melanoma in images is crucial for the early-stage detection of melanoma. In addition, this way the predicted lesion diagnosis were comparable with the verified histopathological diagnosis, as well as, with the earlier study by Pölönen et al. (2019). The histopathological diagnosis contained only one diagnosis per image, therefore it was used as the ground truth class in the evaluation of the lesion classification. To conclude, the evaluation of the lesion classification used the same metrics that we introduced earlier – precision, recall, f1-score, and specificity. As mentioned previously, the metrics of precision, recall, and specificity have been widely used in medical diagnosis tasks, but also they are the most used metrics in the HSI studies, especially in melanoma classification, and enabled the comparison of these studies (Johansen et al. 2020; Pölönen et al. 2019).

Risk class	Lesion type	Type
1	Malignant Melanoma	Very dangerous
2	Lentigo Maligna	Not very dangerous
3	Dysplastic Nevus	Increased risk
4	Benign Nevus	Not dangerous

Table 4. The risk classes for the lesion types used in this study. Risk class one presents the most severe lesion type and risk class four presents not dangerous lesion type.

4 Results

In the previous chapter, the materials and methods of this thesis were introduced. In this chapter, the results of the implemented Unet model and the implemented ResNet-Unet model are summarized. First, we will begin by examining the results of the semantic segmentation, which are presented in Section 4.1. Then, in Section 4.2 we will continue by reporting the lesion classification results.

4.1 Semantic segmentation results

In this section the results of the models performance on semantic segmentation are reported. The models have been evaluated by using the four metrics – precision, recall, specificity, and f1-score that were introduced in Section 3.3. Next, the results are first compared between the two implemented models, and then the segmentation results of each model are presented individually in Section 4.1.1 and in Section 4.1.2.

Table 5 presents the mean and standard deviation results of the semantic segmentation with the two implemented models – Unet, and ResNet-Unet. The mean of the results was calculated from the test set results over the 5-fold cross-validation. As can be seen from Table 5, the implemented ResNet-Unet architecture performed better on semantic segmentation with respect to several metrics. With precision metric, the implemented ResNet-Unet performed slightly better, with a difference of 0.11 percentage points to the implemented Unet model's performance. In addition, the implemented ResNet-Unet performed best with recall metric, where the difference between the models was 0.33 percentage points. F1-score for the implemented ResNet-Unet was 0.21 percentage points better than the performance of the implemented Unet model. On the other hand, the specificity evaluation metric did not show any difference in the performance between the two model types.

4.1.1 Semantic segmentation results of the implemented U-net model

Next, we will inspect visually the semantic segmentation results of the implemented Unet model. First, we will inspect the segmentation predictions with least misclassified pixels,

Model	Precision (%)		Recall (%)		Specificity (%)		F1-score (%)	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Unet	92.61	1.65	91.73	1.80	98.79	0.26	92.17	1.72
ResNet-Unet	92.72	1.28	92.06	1.48	98.79	0.21	92.38	1.37

Table 5. The summary of the segmentation results between the two implemented models. The table shows the mean and standard deviation of the experimental results on both models over the 5-fold cross-validation.

and then we will inspect the segmentation predictions with most misclassified pixels. Figure 19 shows four images of lesions with only few misclassified pixels in segmentation results obtained from the implemented Unet model. Unfortunately, the model only obtained a few high-quality segmentations. Therefore, in Figure 19 we have presented only two different types of lesions: malignant melanoma, and dysplastic nevus, as the model was not able to obtain good segmentation results with benign nevus or lentigo maligna.

Figure 20 presents the low-quality segmentation results of the implemented Unet model. This figure contains five lesion types, one lesion type per row. Similarly to Figure 19, first image in a row shows the hyperspectral false color image in RGB, next the annotated image, and finally the predicted segmentation of the model. In Figure 20 all of the lesion types are presented: benign nevus, dysplastic nevus, lentigo maligna, and malignant melanoma. In Figure 20 the rows one, two, and four contain segmentation predictions with multiple different lesion types in an image, whereas the annotated image always contained one lesion type – the verified histopathological result. In addition, the prediction of the normal skin area and imaging artefact areas (referred as none class) were widely misclassified by the implemented Unet model, as can be seen in Figure 20 on rows three and four. Therefore, the segmentation of borders were not specific in all predictions. In many cases the lesions, especially malignant melanomas, were falsely segmented to contain marker or none classes. Few of such examples are seen in the figure on rows two, three, and five.

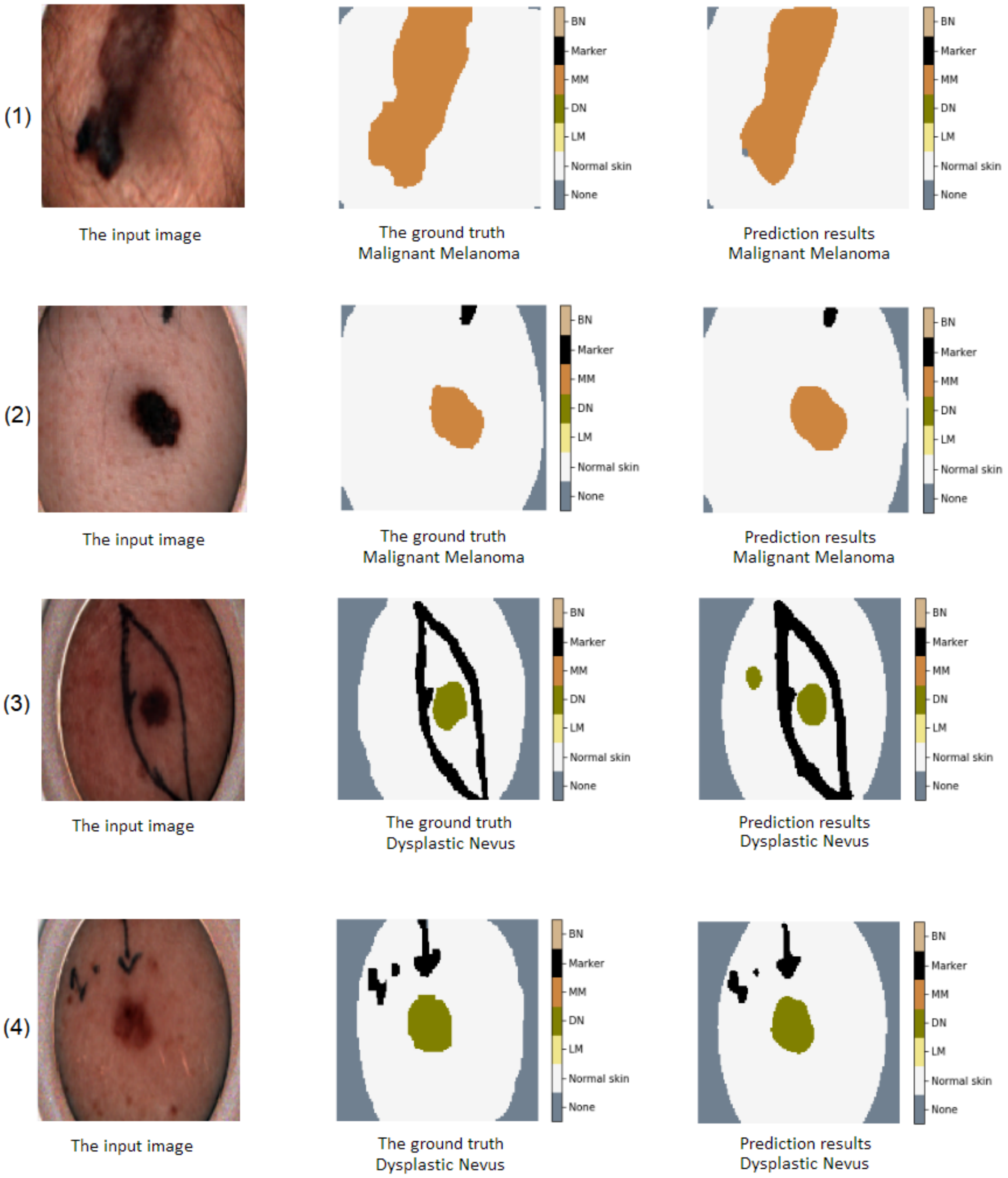

Figure 19. High-quality semantic segmentation results of lesions by the implemented Unet model. Each row presents one lesion: first image per row is the hyperspectral false color image in RGB, the second image presents the annotated image, and the last image presents the segmentation results of the implemented Unet model. The class labels for the segmentations were: bening nevus (BN), dysplastic nevus (DN), lentigo maligna (LM), malignant melanoma (MM), marker, none, and normal skin.

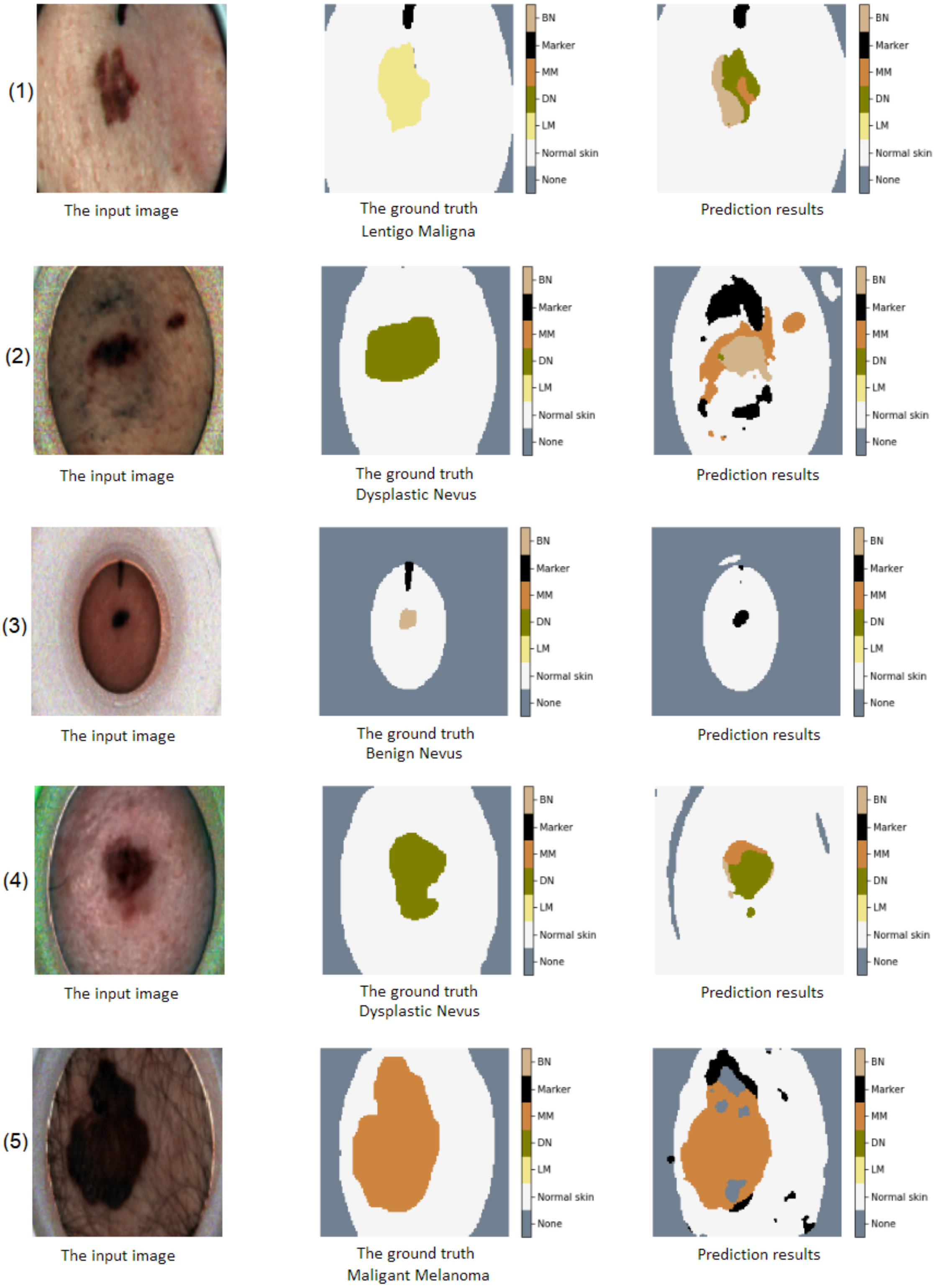

Figure 20. Examples of low-quality semantic segmentation results of lesions by the implemented Unet model. The class labels for the segmentations were: BN, DN, LM, MM, marker, none, and normal skin.

4.1.2 Semantic segmentation results of the implemented ResNet-Unet model

Next, we will visually inspect the segmentation prediction results of the implemented ResNet-Unet model. First, in Figure 21 we present the high-quality segmentation predictions with least falsely predicted pixels, and then in Figure 22 we present the segmentation predictions that contained the most misclassified pixels.

Figure 21 shows the high-quality segmentation predictions for three of the lesion types: malignant melanoma, lentigo maligna, and dysplatic nevus. Similarly to Figure 19 and Figure 20, each row has one lesion type that contains three images – the hyperspectral false color image in RGB, the annotated classes, and the segmentation prediction result. The segmentation predictions of benign nevus typically had some falsely segmented pixels, thus, we excluded those from the well segmented cases. As can be seen from Figure 21, the model had almost no falsely segmented pixels in these predictions. Interestingly, the misclassified pixels seem to show higher quality segmentations of the lesions than the provided annotation images present.

On the contrary to the implemented Unet model, the implemented ResNet-Unet model had only few low-quality segmentations of lesions. Figure 22 shows five examples of these low-quality segmentation predictions, from the lesion types of benign nevus, dysplastic nevus, and malignant melanoma. The predictions of lentigo maligna did not have great misclassifications, therefore we excluded it from the low-quality segmentation presentation. We can see in Figure 22 that with rows one, three, and four the image border predictions were misclassified, especially with pixels of normal skin and imaging artefact (referred as the none class). In addition, some images contained more than one lesion type per prediction, whereas the annotated image always contained one histopathologically verified lesion type, as can be seen in figure rows one, two, and five. Furthermore, the figure rows one, two, three, and five shows the misclassifications of the marker class. To conclude, the model had fewer false segmented predictions, but still the lesion areas and borders in the predictions were not always segmented correctly.

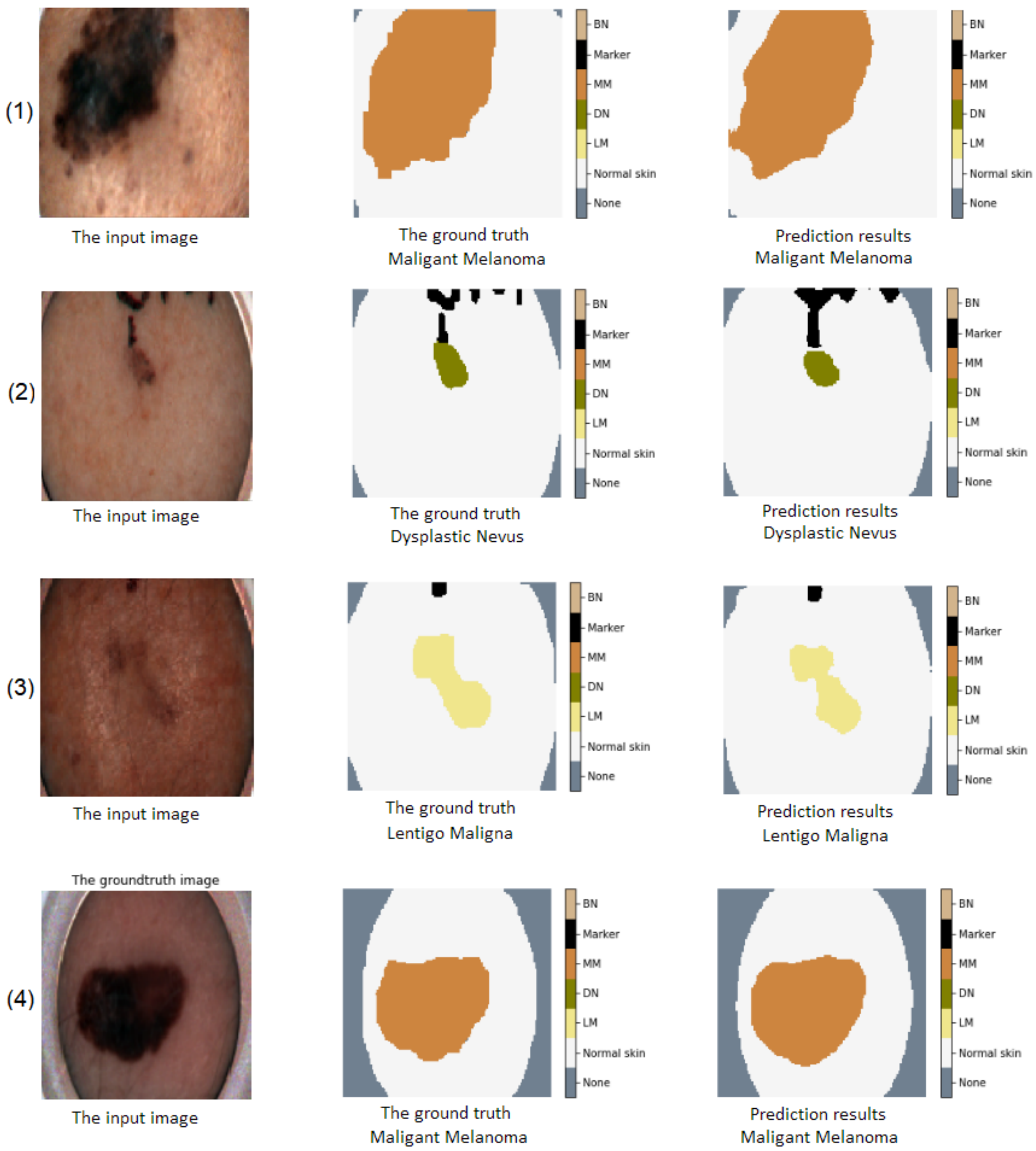

Figure 21. Examples of high-quality semantic segmentation results of lesions by the implemented ResNet-Unet model. Each row of the figure presents one lesion type in three forms: first as a HSI false color RGB image, then as an annotated image, and lastly as a segmentation prediction. The class labels for the segmentations were: bening nevus (BN), dysplastic nevus (DN), lentigo maligna (LM), malignant melanoma (MM), marker, none, and normal skin.

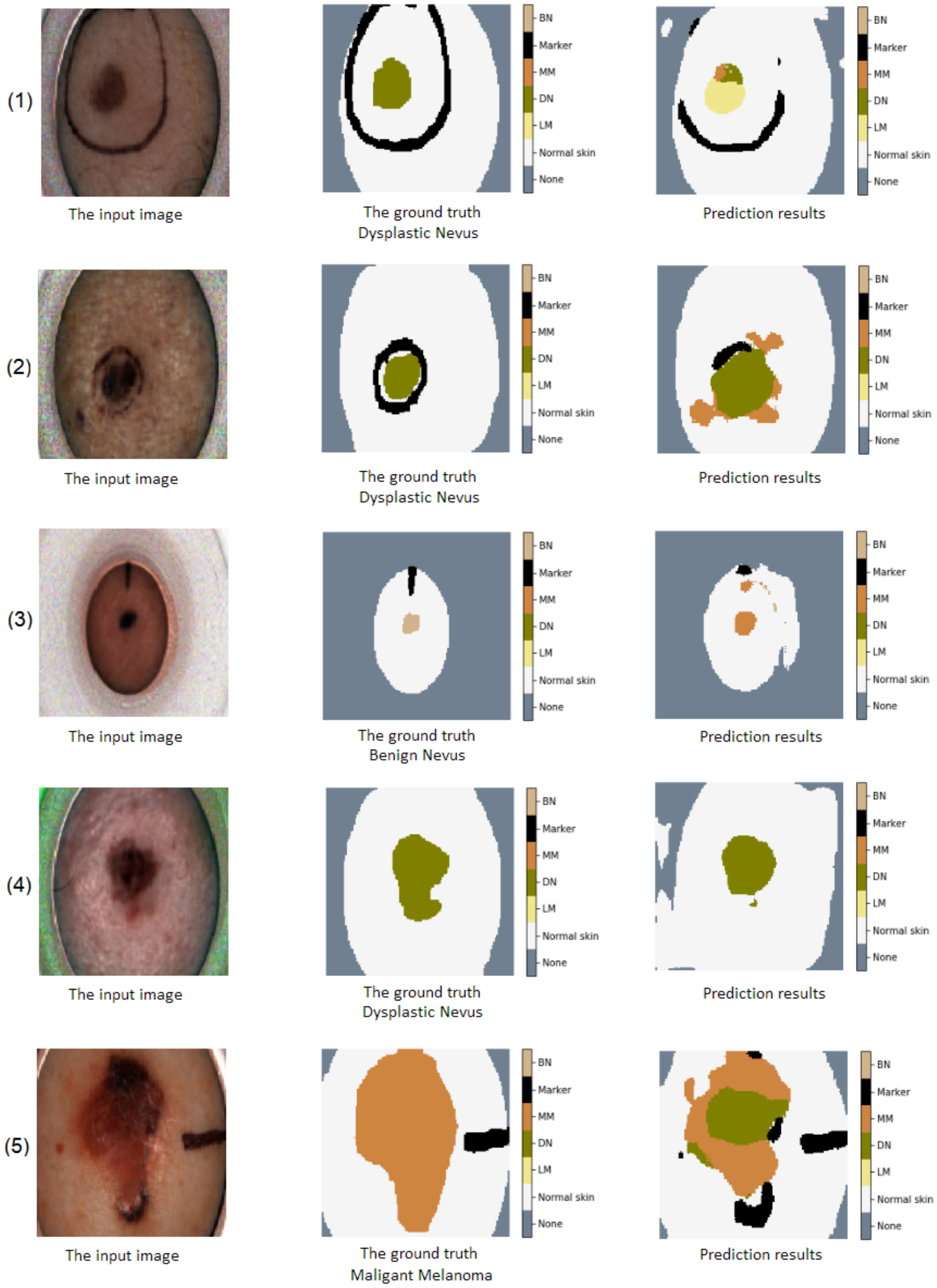

Figure 22. Examples of the implemented ResNet-Unet model results with low-quality semantic segmentations of lesions. The class labels for the segmentations were: BN, DN, LM, MM, marker, none, and normal skin.

4.2 Lesion classification

The previous section presented the segmentation results for the whole image. We will now continue to examine the lesion classification results. First, we will present the overall classification results of both models concerning all the lesion types. Then, we will compare the results of the models by examining each lesion type classification result individually.

Figure 23 shows two confusion matrices, one for the implemented Unet model and the other for the implemented Resnet-Unet model. These matrices report all the predictions of lesion types BN, DN, LM, and MM over the 5-fold cross-validation. The matrices on the figure contain the predicted label on the *x*-axis and the true label on the *y*-axis. The diagonal axis presents all of the true positive predictions for each lesion type, and the off-diagonal items report the falsely classified lesion types. The true positives were the correct predictions for a specific lesion type, when the histopathological analysis ground truth labels and the predicted labels contained the same class. From Figure 23 we can see that the implemented Unet model had two images of lesions where it was not able to predict any lesion type, whereas the implemented ResNet-Unet model was able to predict for all of the images. Therefore, there is a none class label only in the confusion matrix of the implemented Unet model. From the confusion matrices it can be seen that the implemented Unet model had two false negatives for malignant melanoma, where as the implemented ResNet-Unet model had only one false negative in MM classification. On the contrary, the lesion type LM had similar results with both of the models, three false negatives. Furthermore, the DN with the implemented Unet model predictions contained 19 false negatives, when the implemented ResNet-Unet model predicted 15 false negatives for DN. On the other hand, the implemented Unet model predicted all the 14 BN cases as false negatives, and the implemented ResNet-Unet model predicted 10 false negatives for BN. From Figure 23 it can be concluded that the implemented ResNet-Unet had better results with overall lesion type prediction as well as in melanoma classification.

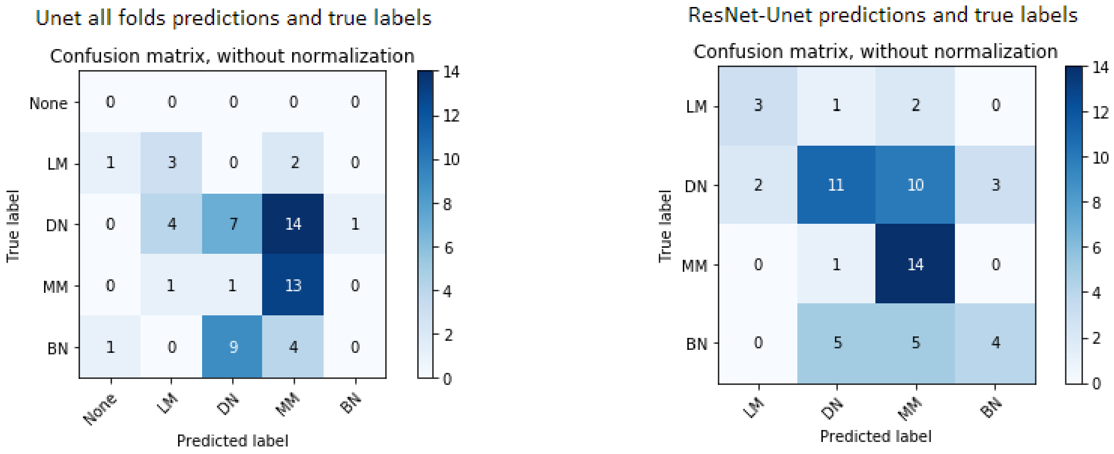

Figure 23. The confusion matrix shows the predicted label and the true label for both of the implemented models with all the lesion types over the 5-fold cross-validation. The diagonal axis presents all the true positive predictions for each lesion type, and the off-diagonal items report the falsely classified lesion types. The class labels for the lesion types are: lentigo maligna (LM), dysplastic nevus (DN), malignant melanoma (MM), and bening nevus (BN). The implemented Unet model was not able to classify all lesion types, therefore a none label is only included in its confusion matrix.

As we have now seen the overall lesion classification results with all the lesion types, we will now continue to report the classification results individually for each lesion type. First we will inspect the results for malignant melanoma, next for lentigo maligna, then for dysplastic nevus, and finally for benign nevus. Table 6 presents the malignant melanoma results for both of the implemented models. As can be seen from the table, the implemented ResNet-Unet model was able to predict malignant melanoma better in all of the metrics – precision by 13.55 percentage points better, recall by 6.66 percentage points better, specificity by 7.75 percentage points better, and f1-score by 8.73 percentage points better. However, the variation of the results computed over the 5-fold cross-validation were somewhat large with both models, as seen in the standard deviation results.

In Table 7 we can see the classification results of lentigo maligna for both of the models. The dataset contained only a few examples of LM, which affected the prediction results. Both of the models had the same results on recall metric, but the implemented ResNet-Unet model

Malignant Melanoma								
Model	Precision (%)		Recall (%)		Specificity (%)		F1-score (%)	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Unet	40.95	10.76	86.67	16.33	55.10	19.21	55.00	11.83
ResNet-Unet	54.50	23.36	93.33	13.33	62.85	23.04	63.73	11.36

Table 6. The summary of the malignant melanoma classification on both models. The table shows the malignant melanoma classification mean and standard deviation results of the 5-fold cross-validation.

was capable to perform better on metrics: precision (10.00 percentage points), specificity (5.30 percentage points), and f1-score (3.34 percentage points). As seen in the table, the standard deviation was quite large on both of the models.

Lentigo Maligna								
Model	Precision (%)		Recall (%)		Specificity (%)		F1-score (%)	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Unet	40.00	37.42	50.00	44.72	90.88	5.77	43.33	38.87
ResNet-Unet	50.00	44.72	50.00	44.72	96.18	4.69	46.67	40.00

Table 7. The summary of the lentigo maligna classification on both models. The table shows the lentigo maligna classification mean and standard deviation results of the 5-fold cross-validation.

We can see the results of dysplastic nevus classification by both of the implemented models from Table 8. The table shows the mean and standard deviation results computed over the 5-fold cross-validation. The implemented ResNet-Unet model performed better when inspecting all the evaluation metrics, but the overall predictions of dysplastic nevus were quite inaccurate as we can see from the table.

Dysplastic Nevus								
Model	Precision (%)		Recall (%)		Specificity (%)		F1-score (%)	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Unet	50.00	27.39	26.00	12.00	72.50	17.34	31.27	10.26
ResNet-Unet	57.43	33.69	44.00	32.00	80.36	14.73	44.89	27.77

Table 8. The summary of dysplastic nevus classification on both models. The table shows the dysplastic nevus classification mean and standard deviation results of the 5-fold cross-validation.

Finally, we will inspect the classification results for benign nevus, which are presented in Table 9. It can be seen from the table that the implemented Unet was not able to predict benign nevus specifically, therefore precision, recall, and f1-score metrics have values zero. On the other hand, the specificity for the implemented Unet model is higher than for the implemented ResNet-Unet model (3.69 percentage points), which is quite misleading as the implemented Unet model had difficulties to predict bening nevus, which was also seen in Figure 23. All in all, the implemented ResNet-Unet model was better in predicting benign nevus, nevertheless, the prediction for benign nevus were poor with both of the models.

Benign Nevus								
Model	Precision (%)		Recall (%)		Specificity (%)		F1-score (%)	
	Mean	*Std*	*Mean*	*Std*	*Mean*	*Std*	*Mean*	*Std*
Unet	0.00	0.00	0.00	0.00	98.18	3.64	0.00	0.00
ResNet-Unet	65.00	43.59	30.00	16.33	94.55	10.91	15.71	20.40

Table 9. The summary of bening nevus classification on both models. The table shows the bening nevus classification mean and standard deviation results of the 5-fold cross-validation.

In conclusion, Figure 24 and Figure 25 summarizes the amount of lesion type predictions that were made for each actual lesion type by the implemented models. First, the Figure 24 shows the predictions of the implemented Unet model. As we can see from the figure, malignant melanoma was mostly predicted correctly, but few misclassifications also occurred. Malignant melanoma was falsely predicted once as lentigo maligna and once as dysplastic nevus. Lentigo maligna was predicted three times correctly, but three times it was also predicted falsely. Lentigo maligna was mostly misclassified as malignant melanoma. Dysplastic nevus and benign nevus were almost always misclassified instead of being correctly classified. Dysplastic nevus was mostly falsely predicted as malignant melanoma, whereas bening nevus was mostly falsely classified as dysplastic nevus.

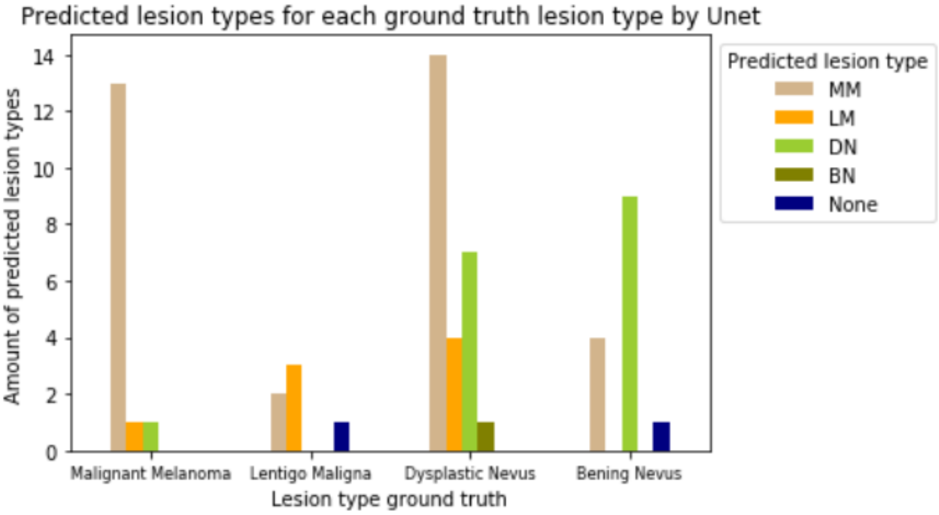

Figure 24. The overall results for predicting each lesion type by the implemented Unet model. The figure shows the graph for each actual lesion type in the *x*-axis, and the predictions the model predicted for the specific lesion type in the *y*-axis.

Finally, the lesion predictions of the implemented ResNet-Unet model are seen in Figure 25. We can see from the figure that malignant melanoma was nearly always classified correctly. Only one time the actual malignant melanoma was falsely predicted as dysplastic nevus. On the other hand, the rest of the lesion types were not classified with great success as we can see from the figure. Lentigo maligna had same amount of misclassifications as it had correct classifications. Lentigo maligna was mostly misclassified as malignant melanoma. Dysplastic nevus was mostly misclassified as malignant melanoma. Lastly, bening nevus was mostly misclassified as malignant melanoma but also as dysplastic nevus.

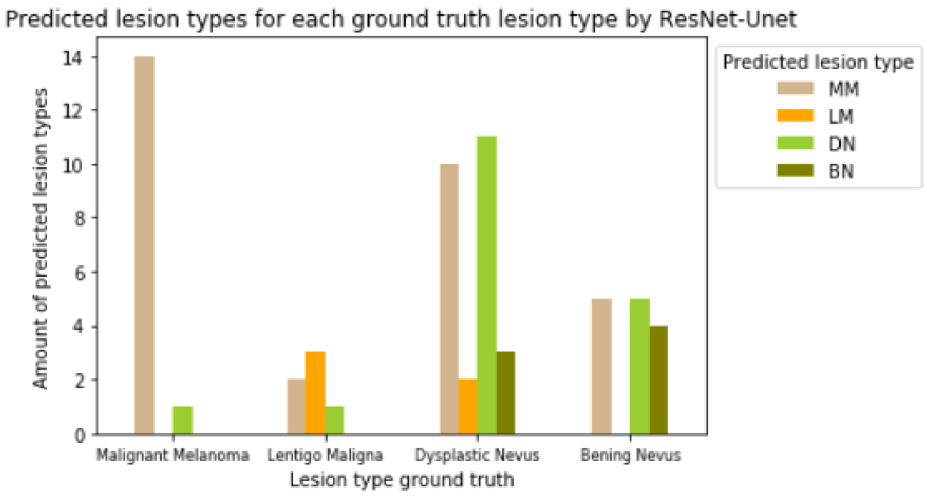

Figure 25. The overall results for predicting each lesion type by the implemented ResNet-Unet model. The figure shows the graph for each actual lesion type in the *x*-axis, and the predictions the models predicted for the specific lesion type in the *y*-axis.

5 Discussion

In the previous chapter the results of this study were described. In this chapter we will discuss the results and the limitations of the study in more detail. Finally, the suggestions for future research are presented in Section 5.1.

With respect to the first research question, it was found that the implemented ResNet-Unet model was able to segmentate the hyperspectral images slightly better than the implemented Unet model, as we saw in Table 5 in Section 4.1. However, when the segmentation predictions were visually inspected in Section 4.1.1 and in Section 4.1.2, the difference between the predictions of the models was quite significant. A source of uncertainty in the segmentation results arise from leaving the imaging covering tube in the corners of the images, as they contained some noise in some of the images. On the other hand, the covering tube was also annotated, and we had no unlabelled pixels in the images. This may have helped the models to segmentate more accurately, when visually inspecting the segmentation results with the segmentation results of the earlier study by Pölönen et al. (2019). Another issue was that some of the lesion area predictions contained multiple lesion types. Further investigation is needed to verify this finding histopathologically. In addition, a major limitation of this study was that the annotations were not performed by a professional. In conclusion, further investigation together with domain experts is needed to delineate lesion borders more specifically, and to gain verified ground truth annotations for semantic segmentation.

To our knowledge, the overall segmentation results containing all the annotated labels with novel hyperspectral data of several lesions have not been previously reported by using similar metrics. As the study area of combining HSI with deep learning in lesion segmentation is quite novel, there are less studies in the field (Johansen et al. 2020). It seems that most of the previous semantic segmentation studies have reported their contribution with classification results of specific lesions, which we will compare later in this section. Nevertheless, there have been studies using RGB images in lesion segmentation, but these studies are not entirely comparable with the novel HSI data segmentation results where we used multiple lesion types. It seems that the studies segmenting the RGB image areas were reported by segmenting the whole image area in binary form, for instance to melanoma and to non-melanoma pixels (Alom et al. 2018). As Johansen et al. (2020) observed in their study, the lack of publications of combining HSI lesion data with deep learning may be explained by the lack of a large and publicly available HSI lesion dataset.

With respect to the second research question, this study found that the implemented ResNet-Unet model was notably better in the lesion classification task when compared to the implemented Unet model, as seen in Section 4.2. When inspecting the two figures: Figure 24 and Figure 25 it can be seen that both models were especially good in predicting malignant melanoma, but the predictions of other lesion types were not as successful. This result may be due to the fact that in the lesion classification we used the method of classifying an image by the most dangerous lesion type, as explained in Section 3.3.2. Furthermore, this may be due to the fact that the threshold value of the classifier was not modified in either of the models. Thus, modifying the threshold value of the classifier could improve the overall classification of the model, but it could also increase the false negatives of malignant melanoma. This finding may also support the fact that the lesion types may visually resemble each other (Jerant et al. 2000; Rigel and Carucci 2000), but also the finding that the spectral presentation of these lesions may contain somewhat overlapping distributions (Pölönen et al. 2019). Therefore, the lesion types can be difficult to tell apart. In addition, it seems possible that these results would benefit from a larger and more balanced dataset.

This study has shown that of the two architecture types we implemented, the implemented ResNet-Unet model performed better than the implemented Unet model. These results are likely to be related to the implemented ResNet-Unet being a deeper model and using residual connections between the fully convolutional layers. This further supports the idea that deeper models enhance the predictions (Krizhevsky et al. 2012; Simonyan and Zisserman 2014), but also the fact that fully convolutional neural networks can improve segmentation results (Long et al. 2015). It is also possible that the usage of residual connections improved the results, as previous study has shown (He et al. 2015a). In accordance with the present results, previous studies have demonstrated that combining the structure of the Unet model with residual connections help the model to converge faster and improve the results (Milletari et al. 2016; Zhang et al. 2018).

The third question in this research was to determine whether we can improve the lesion classification when compared to the study by Pölönen et al. (2019). We compared the implemented ResNet-Unet model with the results of the CNN 2D model by Pölönen et al. (2019), as our models only used the 2D convolution operations. From Table 10 we can see the results between the models with respect to all of the lesion types. One interesting finding was that we were able to improve the classification of LM significantly as we can see from Table 10. Regarding to DN and MM classification we were able to improve them according to two of our metrics, but one of our metrics was worse than in the previous study. In BN classification we were only able to improve one metric, but two metrics remained better in the previous study. These differences may be explained by using less synthetic data in the training phase, and by using smaller dimensional images in our models. Moreover, the possible interference of using whole image augmentations instead of using pixel-wise augmentations cannot be ruled out.

Lesion type	Model	Precision	Recall	Specificity	F1-score
Malignant Melanoma	2D CNN	35.00	**100.00**	12.00	-
	ResNet-Unet	**54.50**	99.33	**62.85**	63.73
Lentigo Maligna	2D CNN	9.00	17.00	64.00	-
	ResNet-Unet	**50.00**	**50.00**	**96.18**	46.67
Dysplastic Nevus	2D CNN	33.00	8.00	**81.00**	-
	ResNet-Unet	**57.43**	**44.00**	80.36	44.89
Bening Nevus	2D CNN	**100.00**	7.00	**100.00**	-
	ResNet-Unet	65.00	**30.00**	94.55	15.71

Table 10. Comparison of the implemented ResNet-Unet model and the 2D CNN model (Pölönen et al. 2019) results with all of the lesion types. The dashes in the table indicate that the results were not reported with such metric.

In addition, we compared the results of the implemented ResNet-Unet model in melanoma classification with other studies, which had used different HSI lesion data. Compared to a study by Hosking et al. (2019), they had better result on metric recall (6.67 percentage points), however our result in specificity (26.85 pp.) was significantly better. When compared to a study by Gu et al. (2018) we were able to get better result with recall (9.33 pp.), but their result in specificity (9.25 pp.) was better. The comparison between these studies indicated that our results had a few improvements. The generalisability of the results are limited with the facts that the studies by Hosking et al. (2019) and Gu et al. (2018) used different datasets, different amounts of data, different lesion types, and different methods as discussed in Section 2.4.1. Therefore, the comparison offers only general trends.

Although the current study is based on a small dataset, the findings are promising. The findings of our study with melanoma classification contained a few improvements. As indicated by Johansen et al. (2020) it is important to gain high results with recall in melanoma detection, but gaining reasonable specificity together with high recall has previously been hard. This we were able to improve. In addition, our results support the earlier finding by Pölönen et al. (2019) that the usage of deep learning with HSI lesion data can outperform the classification of MM when compared to clinical diagnoses by Heal et al. (2008). Moreover, we succeeded to improve, according to few of the metrics, the classification of other lesion types as well. As Johansen et al. (2020) implicated, the classification of non-melanoma lesions is also important, as these lesions are hard to distinguish by general practitioners. Our findings are a small step towards improving the delineation and classification of several different lesion types, nevertheless further work is required.

These findings should be interpreted with caution due to some limitations as indicated earlier. Notable is also the fact that the models were trained only once with the 5-fold cross-validation. Further repetitions are needed to minimize variation in the results. Moreover, it is worth noticing that the data was not diverse, as it contained patients from two cities in Finland. In spite of the limitations, this study certainly adds value to use deep semantic segmentation models with novel hyperspectral image data, to ensure we can improve the early detection of skin cancer and improve to delineate the tumour borders more accurately in clinical examinations in the future.

5.1 Suggestions for further research

Despite the promising results, future investigations are needed to improve the lesion classification and the overall segmentation results. A natural progression of this work could be to annotate the images with professional help. Comparison by using HSI data and RGB data with deep learning models could be investigated in the future. As there have been less studies about pixel level approaches in lesion segmentation according to Johansen et al. (2020), this method could be compared with the whole image segmentation. Future work could also experiment with different types of augmentation methods. However, if a larger dataset would be available it should be used, as more data should help to improve both the segmentation results and the lesion classification results.

Moreover, we suggest experiments with several different architectures and state-of-the-art networks. For example, future experiments could be done by expanding the Unet model with both residual connections, and with recurrent convolution neural network architecture, which was found to improve the network results in the study by Alom et al. (2018). Furthermore, different convolutions, such as 3D convolutions, could be tested with these models. The study by Çiçek et al. (2016) have evidence that 3D convolutions in Unet model can improve the results of the network. The usage of 3D convolutions and combination of different convolutions was also found useful in the study by Pölönen et al. (2019).

6 Conclusions

The main goals of this study were to compare the two different deep learning architectures for semantic segmentation, and to try to improve the lesion classification results compared to the previous study by Pölönen et al. (2019). We implemented two deep learning models, the U-net model and the ResNet-Unet model. We compared these models by their ability to segmentate lesions and to classify lesions. The data used in this study was novel hyperspectral image dataset of lesions. The major limitation of the dataset was its small size of 61 images, nevertheless the hyperspectral images were quite large – $1920 \times 1200 \times 120$ pixels. The dataset was expanded by using data augmentations while training the deep learning models.

We found that the implemented ResNet-Unet model obtained better results in both of the tasks – in semantic segmentation and in lesion classification. To our knowledge the semantic segmentation results of multiple lesion types and labels have not been presented with the novel HSI data in previous studies, and it seems that this study lays groundwork for future research. Our attempt to improve melanoma classification results were minor, nevertheless we were able to improve drastically the classification of lentigo maligna. In addition, this work was able to slightly improve the classification of dysplastic nevus. The results of our study are consistent with the findings of Pölönen et al. (2019), which found that using deep learning with semantic segmentation to predict melanoma from the novel HSI data can slightly outperform clinical diagnoses of melanoma classification. Overall, this study strengthens the idea that deep neural networks are able to learn highly complicated features, and when combined with HSI data they might have potential to help to improve melanoma detection and help to delineate lesion borders.

Bibliography

Adam, Elhadi, Onisimo Mutanga, and Denis Rugege. 2010. “Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review”. *Wetlands Ecology and Management* 18 (3): 281–296.

Alexandrescu, D. T. 2009. “Melanoma costs: a dynamic model comparing estimated overall costs of various clinical stages”. *Dermatology online journal* 15 (11).

Alom, Md Zahangir, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. 2018. *Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation.* arXiv: `1802.06955 [cs.CV]`.

Arganda-Carreras, Ignacio, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. 2015. “Crowdsourcing the creation of image segmentation algorithms for connectomics”. *Frontiers in neuroanatomy* 9:142.

Argenziano, G., L. Cerroni, I. Zalaudek, S. Staibano, R. Hofmann-Wellenhof, N. Arpaia, R. M. Bakos, et al. 2012. “Accuracy in melanoma detection: a 10-year multicenter survey”. *Journal of the American Academy of Dermatology* 67 (1): 54–59.

Badrinarayanan, Vijay, Ankur Handa, and Roberto Cipolla. 2015. “Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling”. *arXiv preprint arXiv:1505.07293.*

Bishop, Christopher M. 2006. *Pattern recognition and machine learning.* Information science and statistics. Softcover published in 2016. New York, NY: Springer. `https://cds.cern.ch/record/998831`.

Boggs, Thomas. 2014. *Spectral Python, free open source software GNU licenced, Python module for Hyperspectral Data.* Version 0.20. `http://www.spectralpython.net/`.

Bottou, Léon. 2010. “Large-scale machine learning with stochastic gradient descent”. In *Proceedings of COMPSTAT’2010,* 177–186. Springer.

Braun, R. P., H. S. Rabinovitz, M. O., A. W. Kopf, and J.H. Saurat. 2005. “Dermoscopy of pigmented skin lesions”. *Journal of the American Academy of Dermatology* 52 (1): 109–121.

Carrasco, Oscar, Richard B Gomez, Arun Chainani, and William E Roper. 2003. “Hyperspectral imaging applied to medical diagnoses and food safety”. In *Geo-Spatial and Temporal Image and Data Exploitation III,* 5097:215–221. International Society for Optics and Photonics.

Chang, Chein-I. 2007. *Hyperspectral data exploitation: theory and applications.* John Wiley & Sons.

Chollet, François, et al. 2015. *Keras.* `https://keras.io`.

Choromanska, Anna, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. 2014. *The Loss Surfaces of Multilayer Networks.* arXiv: `1412.0233 [cs.LG]`.

Çiçek, Özgün, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. “3D U-Net: learning dense volumetric segmentation from sparse annotation”. In *International conference on medical image computing and computer-assisted intervention,* 424–432. Springer.

Ciresan, D., A. Giusti, L. M. Gambardella, and J. Schmidhuber. 2012. “Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images”. In *Advances in Neural Information Processing Systems 25,* edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 2843–2851. Curran Associates, Inc.

Cummins, Deborah L, Jordan M Cummins, Hardin Pantle, Michael A Silverman, Aimee L Leonard, and Arjun Chanmugam. 2006. “Cutaneous malignant melanoma”. In *Mayo clinic proceedings,* 81:500–507. 4. Elsevier.

Drozdzal, Michal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. 2016. *The Importance of Skip Connections in Biomedical Image Segmentation.* arXiv: `1608.04117 [cs.CV]`.

Dumoulin, Vincent, and Francesco Visin. 2016. “A guide to convolution arithmetic for deep learning”. *ArXiv e-prints* (). eprint: `1603.07285`.

Eriksson, Thérèse, and Gustav Tinghög. 2015. “Societal cost of skin cancer in Sweden in 2011”. *Acta dermato-venereologica* 95 (3): 347–348.

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. 2017. “Dermatologist-level classification of skin cancer with deep neural networks”. *Nature* 542:115–118.

Fabelo, Himar, Verónica Melián, Beatriz Martínez, Patricia Beltrán, Samuel Ortega, Margarita Marrero, Gustavo M Callicó, Roberto Sarmiento, Irene Castaño, Gregorio Carretero, et al. 2019. “Dermatologic Hyperspectral Imaging System for Skin Cancer Diagnosis Assistance”. In *2019 XXXIV Conference on Design of Circuits and Integrated Systems (DCIS),* 1–6. IEEE.

Feng, Yao-Ze, and Da-Wen Sun. 2012. “Application of hyperspectral imaging in food safety inspection and control: a review”. *Critical reviews in food science and nutrition* 52 (11): 1039–1058.

Fink, C., and H. A. Haenssle. 2017. “Non-invasive tools for the diagnosis of cutaneous melanoma”. *Skin Research and Technology* 23 (3): 261–271.

Friedman, Robert J, Darrell S Rigel, and Alfred W Kopf. 1985. “Early detection of malignant melanoma: the role of physician examination and self-examination of the skin”. *CA: a cancer journal for clinicians* 35 (3): 130–151.

Fukushima, Kunihiko. 1980. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. *Biological cybernetics* 36 (4): 193–202.

Gandini, S., F. Sera, M. S. Cattaruzza, P. Pasquini, D. Abeni, P. Boyle, and C. F. Melchi. 2005a. “Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi”. *European journal of cancer* 41 (1): 28–44.

Gandini, S., F. Sera, M. S. Cattaruzza, P. Pasquini, O. Picconi, P. Boyle, and C. F. Melchi. 2005b. “Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure”. *European Journal of Cancer* 41 (1): 45–60.

Garini, Y., I. T. Young, and G. McNamara. 2006. "Spectral imaging: principles and applications". *Cytometry Part A: The Journal of the International Society for Analytical Cytology* 69 (8): 735–747.

Glorot, Xavier, and Y. Bengio. 2010. "Understanding the difficulty of training deep feedforward neural networks". *Journal of Machine Learning Research - Proceedings Track* 9 (): 249–256.

Gloster Jr, Hugh M, and Kenneth Neal. 2006. "Skin cancer in skin of color". *Journal of the American Academy of Dermatology* 55 (5): 741–760.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press.

Gorriz, Marc, Axel Carlier, Emmanuel Faure, and Xavier Giro-i-Nieto. 2017. *Cost-Effective Active Learning for Melanoma Segmentation.* arXiv: `1711.09168 [cs.CV]`.

Govender, Megandhren, K Chetty, and Hartley Bulcock. 2007. "A review of hyperspectral remote sensing and its application in vegetation and water resource studies". *Water Sa* 33 (2).

Greene, M. H., W. H. Clark, M. A. Tucker, K. H. Kraemer, D. E. Elder, and M. C. Fraser. 1985. "High Risk of Malignant Melanoma in Melanoma-Prone Families with Dysplastic Nevi". *Annals of Internal Medicine* 102 (4): 458–465.

Gu, Yanyang, Yi-Ping Partridge, and Jun Zhou. 2018. "A hyperspectral dermoscopy dataset for melanoma detection". In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis,* 268–276. Springer.

Guy, G. P., and D. U. Ekwueme. 2011. "Years of Potential Life Lost and Indirect Costs of Melanoma and Non-Melanoma Skin Cancer". *PharmacoEconomics* 29 (10): 863–874.

Hall, H Irene, Donald R Miller, Joseph D Rogers, and Barbara Bewerse. 1999. "Update on the incidence and mortality from melanoma in the United States". *Journal of the American Academy of Dermatology* 40 (1): 35–42.

Hariharan, Parameswaran. 2010. *Basics of interferometry.* Elsevier.

Harvey, David. 2011. “Analytical Chemistry 2.0—an open-access digital textbook”. *Analytical and bioanalytical chemistry* 399 (1): 543–666.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015a. *Deep Residual Learning for Image Recognition.* arXiv: `1512.03385 [cs.CV]`.

———. 2015b. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. *CoRR* abs/1502.01852. arXiv: `1502.01852`. `http://arxiv.org/abs/1502.01852`.

Heal, C. F., B. A. Raasch, P. G. Buettner, and D. Weedon. 2008. “Accuracy of clinical diagnosis of skin lesions”. *British Journal of Dermatology* 159 (3): 661–668.

Henry, Corentin, Seyed Majid Azimi, and Nina Merkle. 2018. “Road segmentation in SAR satellite images with deep fully convolutional neural networks”. *IEEE Geoscience and Remote Sensing Letters* 15 (12): 1867–1871.

Hosking, Anna-Marie, Brandon J Coakley, Dorothy Chang, Faezeh Talebi-Liasi, Samantha Lish, Sung Won Lee, Amanda M Zong, Ian Moore, James Browning, Steven L Jacques, et al. 2019. “Hyperspectral imaging in automated digital dermoscopy screening for melanoma”. *Lasers in surgery and medicine* 51 (3): 214–222.

Hubel, D. H., and T. N. Wiesel. 1962. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. *The Journal of physiology* 160 (1): 106–154.

Ioffe, Sergey, and Christian Szegedy. 2015. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* arXiv: `1502.03167 [cs.LG]`.

Jaderberg, Max, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. *Deep Structured Output Learning for Unconstrained Text Recognition.* arXiv: `1412.5903 [cs.CV]`.

Jerant, Anthony F, Jennifer T Johnson, Catherine Demastes Sheridan, and Timothy J Caffrey. 2000. “Early detection and treatment of skin cancer”. *American family physician* 62 (2): 357–368.

Johansen, Thomas Haugland, Kajsa Møllersen, Samuel Ortega, Himar Fabelo, Aday Garcia, Gustavo M Callico, and Fred Godtliebsen. 2020. “Recent advances in hyperspectral imaging for melanoma detection”. *Wiley Interdisciplinary Reviews: Computational Statistics* 12 (1): e1465.

Jones, Hamlyn G, and Robin A Vaughan. 2010. *Remote sensing of vegetation: principles, techniques, and applications.* Oxford university press.

Kai Wang, B. Babenko, and S. Belongie. 2011. “End-to-end scene text recognition”. In *2011 International Conference on Computer Vision,* 1457–1464.

Kingma, Diederik P, and Jimmy Ba. 2014. “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980.*

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

Lasithiotakis, Konstantinos G, Ulrike Leiter, Roman Gorkievicz, Thomas Eigentler, Helmut Breuninger, Gisela Metzler, Waltraud Strobel, and Claus Garbe. 2006. “The incidence and mortality of cutaneous melanoma in Southern Germany: trends by anatomic site and pathologic characteristics, 1976 to 2003”. *Cancer* 107 (6): 1331–1339.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. “Backpropagation Applied to Handwritten Zip Code Recognition”. *Neural Computation* 1 (4): 541–551.

Liu, Z., H. Wang, and Q. Li. 2011. “Tongue Tumor Detection in Medical Hyperspectral Images”. *Sensors* 12 (1): 162–174.

Long, J., E. Shelhamer, and T. Darrell. 2015. “Fully convolutional networks for semantic segmentation”. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 3431–3440.

Lu, G., and B. Fei. 2014. “Medical hyperspectral imaging: a review”. *Journal of Biomedical Optics* 19 (1): 1–24.

Marghoob, Ashfaq A, Lucinda D Swindle, Claudia ZM Moricz, Fitzgeraldo A Sanchez Negron, Bill Slue, Allan C Halpern, and Alfred W Kopf. 2003. “Instruments and new technologies for the in vivo diagnosis of melanoma”. *Journal of the American Academy of Dermatology* 49 (5): 777–797.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2015. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. `https://www.tensorflow.org/`.

McCulloch, W. S., and W. Pitts. 1943. “A logical calculus of the ideas immanent in nervous activity”. *The bulletin of mathematical biophysics* 5 (4): 115–133.

Meyer, A., N. O. Salscheider, P. F. Orzechowski, and C. Stiller. 2018. “Deep Semantic Lane Segmentation for Mapless Driving”. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 869–875.

Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In *2016 fourth international conference on 3D vision (3DV),* 565–571. IEEE.

Neittaanmäki, N., M. Salmivuori, I. Pölönen, L. Jeskanen, A. Ranki, O. Saksela, E. Snellman, and M. Grönroos. 2017. “Hyperspectral imaging in detecting dermal invasion in lentigo maligna melanoma”. *British Journal of Dermatology* 177 (6): 1742–1744.

Neittaanmäki-Perttu, Noora, Mari Grönroos, Leila Jeskanen, Ilkka Pölönen, Annamari Ranki, Olli Saksela, and Erna Snellman. 2015. “Delineating margins of lentigo maligna using a hyperspectral imaging system”. *Acta dermato-venereologica* 95 (5): 549–552.

Neittaanmäki-Perttu, Noora, Mari Grönroos, Taneli Tani, Ilkka Pölönen, Annamari Ranki, Olli Saksela, and Erna Snellman. 2013. “Detecting field cancerization using a hyperspectral imaging system”. *Lasers in surgery and medicine* 45 (7): 410–417.

Nielsen, M A. 2015. *Neural networks and deep learning.* Volume 25. Determination press.

Offidani, A., O. Simonetti, M. L. Bernardini, A. Alpagut, A. Cellini, and G. Bossi. 2002. “General practitioners’ accuracy in diagnosing skin cancers”. *Dermatology* 205 (2): 127–130.

Oliveira, G. L., W. Burgard, and T. Brox. 2016. “Efficient deep models for monocular road segmentation”. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 4885–4891.

Parikh, Rajul, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. 2008. “Understanding and using sensitivity, specificity and predictive values”. *Indian journal of ophthalmology* 56 (1): 45.

Pölönen, I., S. Rahkonen, L. Annala, and N. Neittaanmäki. 2019. “Convolutional neural networks in skin cancer detection using spatial and spectral domain”, volume 10851.

Rigel, D. S., J. K. Rivers, A. W. Kopf, R. J. Friedman, A. F. Vinokur, E. R. Heilman, and M. Levenstein. 1989. “Dysplastic nevi”. *Cancer* 63 (2): 386–389.

Rigel, Darrell S., and John A. Carucci. 2000. “Malignant melanoma: Prevention, early detection, and treatment in the 21st century”. *CA: A Cancer Journal for Clinicians* 50 (4): 215–236.

Ronneberger, O., P. Fischer, and T. Brox. 2015. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015,* edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, 234–241. Cham: Springer International Publishing.

Rosenblatt, F. 1957. “The Perceptron: a perceiving and recognizing automaton”.

Rumelhart, David E, Geoffrey E Hinton, Ronald J Williams, et al. 1986. “Learning representations by back-propagating errors”. *Nature* 323:533–536.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. “ImageNet Large Scale Visual Recognition Challenge”. *International Journal of Computer Vision (IJCV)* 115 (3): 211–252. doi:10.1007/s11263-015-0816-y.

Saari, H., V.V. Aallos, C. Holmlund, J. Malinen, and J. Mäkynen. 2010. “Handheld hyperspectral imager”. *Next-Generation Spectroscopic Technologies III* 7680.

Sallab, Ahmad EL, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. 2017. “Deep reinforcement learning framework for autonomous driving”. *Electronic Imaging* 2017 (19): 70–76.

Salmivuori, M., N. Neittaanmäki, I. Pölönen, L. Jeskanen, E. Snellman, and M. Grönroos. 2019. “Hyperspectral Imaging System in the Delineation of Ill-defined Basal Cell Carcinomas: A Pilot Study”. *Journal of the European Academy of Dermatology and Venereology* 33 (1): 71–78.

Salzer, R, G Steiner, HH Mantsch, J Mansfield, and EN Lewis. 2000. “Infrared and Raman imaging of biological and biomimetic samples”. *Fresenius’ Journal of Analytical Chemistry* 366 (6-7): 712–726.

Schuler, Rebecca L, Paul E Kish, and Cara A Plese. 2012. “Preliminary observations on the ability of hyperspectral imaging to provide detection and visualization of bloodstain patterns on black fabrics”. *Journal of forensic sciences* 57 (6): 1562–1569.

Siam, Mennatullah, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, and Martin Jagersand. 2018. “Rtseg: Real-time semantic segmentation comparative study”. In *2018 25th IEEE International Conference on Image Processing (ICIP),* 1603–1607. IEEE.

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal. 2019. “Cancer statistics, 2019”. *CA: a cancer journal for clinicians* 69 (1): 7–34.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. “Mastering the game of Go with deep neural networks and tree search”. *Nature* 529:484–489.

Simard, P. Y., D. Steinkraus, and J. C. Platt. 2003. “Best practices for convolutional neural networks applied to visual document analysis”. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* 958–963.

Simonyan, Karen, and Andrew Zisserman. 2014. *Very Deep Convolutional Networks for Large-Scale Image Recognition.* arXiv: `1409.1556 [cs.CV]`.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. *J. Mach. Learn. Res.* 15, number 1 (): 1929–1958. ISSN: 1532-4435.

Stang, Andreas, Eero Pukkala, Risto Sankila, Bengt Söderman, and Timo Hakulinen. 2006. “Time trend analysis of the skin melanoma incidence of Finland from 1953 through 2003 including 16,414 cases”. *International journal of cancer* 119 (2): 380–384.

Stuart, B. 2004. “Infrared spectroscopy: Fundamental and applications”. *Google Scholar.*

Sun, Da-Wen. 2010. *Hyperspectral imaging for food quality analysis and control.* Elsevier.

Taha, Abdel Aziz, and Allan Hanbury. 2015. “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool”. *BMC medical imaging* 15 (1): 29.

Tannous, Z., L. H. Lerner, L. M. Duncan, M. C. Mihm, and T. J. Flotte. 2000. “Progression To Invasive Melanoma From Malignant Melanoma In Situ, Lentigo Maligna Type”. *Human pathology* 31 (6): 705–708.

Tomatis, S., M. Carrara, A. Bono, C. Bartoli, M. Lualdi, G. Tragni, A. Colombo, and R. Marchesini. 2005. “Automated melanoma detection with a novel multispectral imaging system: results of a prospective study”. *Physics in medicine & biology* 50 (8): 1675.

Tsao, H., M. B. Atkins, and A. J. Sober. 2004. “Management of Cutaneous Melanoma”. *New England Journal of Medicine* 351 (10): 998–1012.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Vaughan, M. 1989. *The Fabry-Perot interferometer: history, theory, practice and applications.* CRC press.

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”. *British Journal of Dermatology* 159 (3): 669–676.

Weinstock, Martin A. 2006. “Cutaneous melanoma: public health approach to early detection”. *Dermatologic therapy* 19 (1): 26–31.

Wolfe, William L. 1997. *Introduction to imaging spectrometers.* Volume 25. SPIE Press.

Wong, Sebastien C., Adam Gatt, Victor Stamatescu, and Mark D. McDonnell. 2016. *Understanding data augmentation for classification: when to warp?* arXiv: 1609.08764 [cs.CV].

Xu, Yan, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. *Improved Relation Classification by Deep Recurrent Neural Networks with Data Augmentation.* arXiv: 1601.03651 [cs.CL].

Yu, L., H. Chen, Q. Dou, J. Qin, and P. Heng. 2017. "Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks". *IEEE Transactions on Medical Imaging* 36 (4): 994–1004.

Zhang, Zhengxin, Qingjie Liu, and Yunhong Wang. 2018. "Road Extraction by Deep Residual U-Net". *IEEE Geoscience and Remote Sensing Letters* 15, number 5 (): 749–753. ISSN: 1558-0571. doi:10.1109/lgrs.2018.2802944. http://dx.doi.org/10.1109/LGRS.2018.2802944.

Zheludev, V., I. Pölönen, N. Neittaanmäki-Perttu, A. Averbuch, P. Neittaanmäki, M. Grönroos, and H. Saari. 2015. "Delineation of malignant skin tumors by hyperspectral imaging using diffusion maps dimensionality reduction". *Biomedical Signal Processing and Control* 16:48–60.