DEGREE PROJECT IN MEDICAL ENGINEERING,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, *SWEDEN 2017*

Use of Deep Learning in Detection of Skin Cancer and Prevention of Melanoma

MARIA PAPANASTASIOU

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF TECHNOLOGY AND HEALTH

Use of Deep Learning in Detection of Skin Cancer and Prevention of Melanoma

Användning av Djupt Lärande vid Upptäckt av Hudcancer och Förebyggande av Melanom

---

Maria Papanastasiou

June, 2017

Supervisor: Jadran Bandic
Examiner: Rodrigo Moreno

Acknowledgement

I would like to acknowledge the contribution of Uros Gligovic for the completion of this project who supported me both technically and mentally during the whole period of the master thesis.

Furthermore, I would wish to express my gratitude to my family who have supported and encouraged me through all my big steps.

I would also like to express my sincere gratitude to my supervisor, Mr Jadran Bandic for the continuous support, motivation and immense knowledge.

A very special gratitude goes to Mr Konstantinos Tsokanis for helping and supporting me those two last years.

And finally, last but by no means least, also to Dmitry Grishenkov and the supervision group. It was great to cooperate with you.

Thanks for all your encouragement!!

Abstract

Melanoma is a life threatening type of skin cancer with numerous fatal incidences all over the world. The 5-year survival rate is very high for cases that are diagnosed in early stage. So, early detection of melanoma is of vital importance. Except for several techniques that clinicians apply so as to improve the reliability of detecting melanoma, many automated algorithms and mobile applications have been developed for the same purpose.

In this paper, deep learning model designed from scratch as well as the pretrained models Inception v3 and VGG-16 are used with the aim of developing a reliable tool that can be used for melanoma detection by clinicians and individual users. Dermatologists who use dermoscopes can take advantage of the algorithms trained on dermoscopical images and acquire a confirmation about their diagnosis. On the other hand, the models trained on clinical images can be used on mobile applications, since a cell phone camera takes images similar to them.

The results using Inception v3 model for dermoscopical images achieved accuracy 91.4%, sensitivity 87.8% and specificity 92.3%. For clinical images, the VGG-16 model achieved accuracy 86.3%, sensitivity 84.5% and specificity 88.8%. The results are compared to those of clinicians, which shows that the algorithms can be used reliably for the detection of melanoma.

Keywords: Clinical images, Deep Learning, Dermoscopical images, Inception v3, Learning rate, Melanoma, Transfer Learning, VGG-16.

Table of Contents

1 Introduction

Malignant melanoma is the deadliest type of skin cancer. 105.000 new cases of melanomas are estimated annually and 33.000 deaths (Parkin et al., 1999). Australia and New Zealand are in first positions of melanoma incidences, with age-standardized rate 27.9 in men and 25.0 in women (Parkin et al., 1999). As far as Europe is concerned, Scandinavian countries appear the highest incidence rate with Central and Southern countries to follow. This can be attributed in lighter skin type in Scandinavian countries as well as in different recreational activities. The mortality of melanoma incidences in Southern countries is connected mostly to miss opportunities for early diagnosis and limited access to therapy (Siegel et al., 2015).

The 5-year and 10-year relative survival rates for melanoma patients are 92% and 89% respectively (Miller et al., 2016). The survival rate in Australia and New Zealand is approximately 85%. Probably educational campaigns have raised the awareness of people and consequently this has resulted in early diagnosis of melanoma (Parkin et al., 1999).

High survival rates show the importance of early detection of melanomas. Many automated algorithms have been developed so as to offer an assisting tool to clinicians as well as a primary opinion for suspicious skin lesions to individual users. Ramlakhan et al. (2011) suggested a mobile automated skin lesion classification system, which achieved classification by using the features of the lesions and a k-NN (k-nearest neighbor) classifier. The results were not better than a clinician's examination (accuracy: 66.7%, sensitivity: 60.7%, specificity: 80.5%). Another approach was by using Very Deep Residual Networks as proposed by Yu et al. (2016), in which a custom model with more than 50 layers was used for melanoma recognition. Despite the good performance of the algorithm (accuracy: 94.9%, sensitivity: 91.1%, specificity: 95.7%), a lot of time is needed for its training which cannot be achieved during the period of master's project. Also, Ensembles of methods for detection of melanoma have been applied (Nylund, 2016 and Codella et al., 2016) with quite satisfying results (accuracy: 89.3% and 75.5%, sensitivity: 77.1% and 62.7%, specificity: 93.0% and 79.6%, respectively).

Last years, Transfer Learning has gained a lot of popularity for classification tasks because of the good performance as well as the reasonable training time. In this project, Inception v3 and VGG-16 will be implemented with the purpose of choosing the one that provides the best results.

2 Materials

2.1 Datasets

The images that were used for the training and testing of algorithms were acquired from online sources as well as company "Teleskin". The images are dermoscopical (image 1), which were taken with the use of dermoscopy (outlined in Appendix, Chapter 1.2.2) and clinical (image 2) (outlined in Appendix, Chapter 1.2.1).

The dermoscopical images were acquired from ISIC Archive (Isic-archive.com, 2017), PH2 (Mendonça et al., 2013), Dermoscopy Atlas of Menzies (2009) and Teleskin. The diagnosis for all images is confirmed by histopathology examination. Specifically, in Teleskin company ABCD rule of dermoscopy, 7-point checklist and Menzies method are used for diagnosis before surgery and then histopathology. All the dermoscopical images from Teleskin and Menzies dataset are taken in contact with skin, oil was applied in the surface and they are non-polarized. The clinical images that were used, were collected from Dermoscopy Atlas of Menzies (2009) and Teleskin. From the dataset of Teleskin, clinical images include images taken from distance close to the lesion (< 5 cm) as well as images from bigger (5-10 cm) distance. Images from skin lesions on face (lentigo maligna melanoma) and extremities (acral lentiginous melanoma) were excluded. Moreover, images that included only one type of skin lesion were included, while images with both melanoma and nevus were excluded. In the following table, exact number of every dataset is presented.

Table 1: The datasets with the specific number of images of melanomas and nevi.

Dataset	Melanomas	Nevi
ISIC Archive (dermoscopical)	661	967
PH2 (dermoscopical)	40	80
Menzies (dermoscopical)	77	83
Teleskin (dermoscopical)	201	636
Menzies (clinical)	77	83
Teleskin (clinical)	247	450

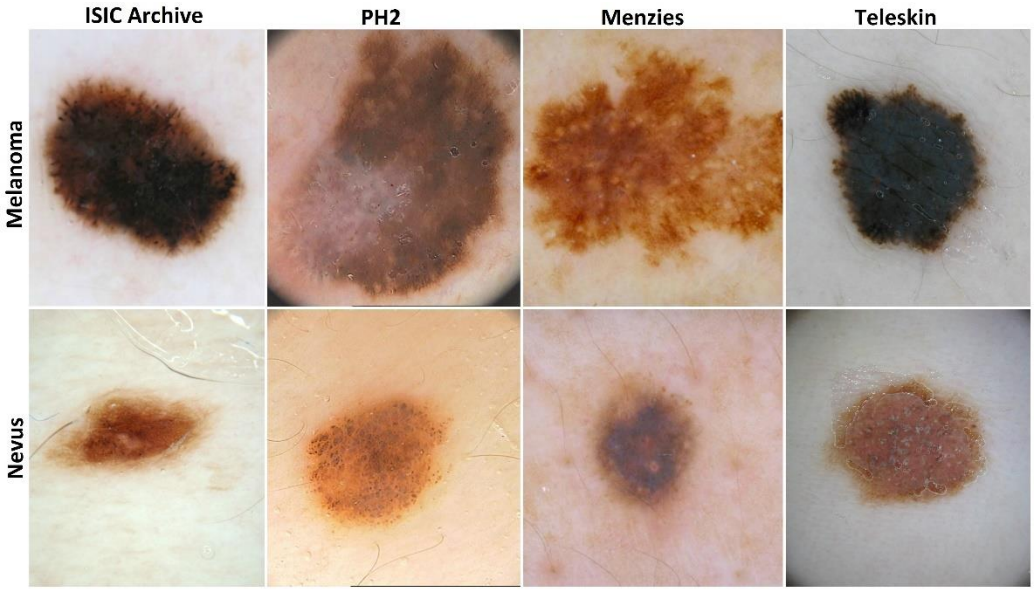

Image 1: Dermoscopical images. First row images of melanomas and second images of nevi. Left images are from ISIC Archive dataset (Isic-archive.com, 2017), second column from PH2 dataset (Mendonça et al., 2013), third from Menzies (Dermoscopy Atlas of Menzies, 2009) and right images from Teleskin.

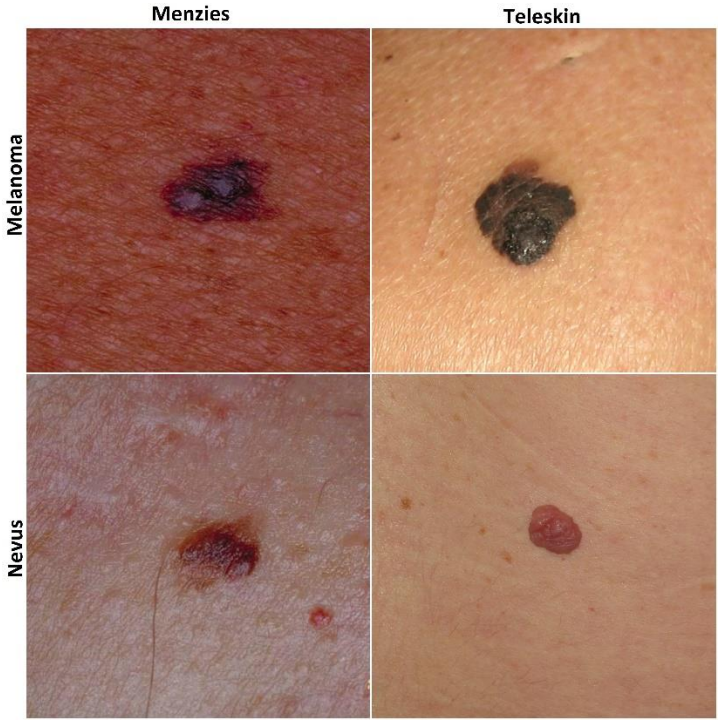

Image 2: Clinical images. Left images from Menzies (Dermoscopy Atlas of Menzies, 2009) and right from Teleskin. In first row, images of melanomas while in second of nevi.

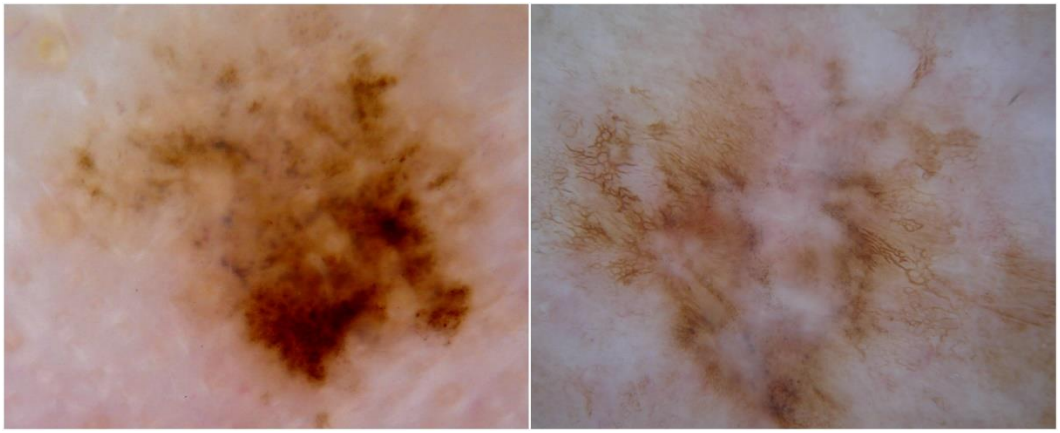

Image 3: Images of melanomas from ISIC Archive dataset (Isic-archive.com, 2017). Left is a full lesion, while the right is a part of a lesion.

From ISIC Archive, two different trainings took place. In the first one, only images of full lesions were used while in the second also images from part of the lesions were included (image 3). For the training and testing, images from only one dataset were used at a time but also mixes of them were formed so as to test the accuracy when different types of images are combined. The combinations that were formed are the followings (Table 2):

Table 2: Datasets that were used for training and testing

Type	Dataset
Dermoscopical	ISIC Archive (full lesions)
	ISIC Archive (all)
	ISIC Archive + PH2
	Teleskin
	Teleskin + Menzies
	Teleskin + ISIC Archive
	Teleskin + Menzies + ISIC Archive + PH2
Clinical	Teleskin
	Teleskin + Menzies

In the above datasets, license for use of the images for educational purposes is provided. Images were resized for each algorithm according to the input demands. More details are found in the "Methodology" chapter.

2.2 Tools

The algorithms are written in Python, and the frameworks that were used were Tensorflow and Keras (outlined in Appendix, Chapter 2.4 and 2.5 respectively).

For the Inception v3 model (Chapter 3.2) because of incompatibility between the versions of the frameworks, the algorithm was trained using a Docker image. Docker is a container technology similar to virtual machines, which consumes less CPU resources and memory. It provides the advantage of independency among different versions of frameworks as well as fast configurations.

The specification of the computer on which trainings and testing took place includes Intel(R) Core(TM) i7-4800MQ CPU @2.70 GHz, 16 GB installed RAM and NVIDIA GeForce GT 920M with 2 GB of memory.

2.3 Hyper-parameters

The hyper-parameters that were used in algorithms and had greater influence in results of the classification are the followings: learning rate, batch size, epoch and number of hidden layers.

Learning rate controls the ability of the algorithm to learn the features of the images during the training (Bengio, 2012). Large learning rate indicates quick adoption of new features. This can be undesirable for the results, since the algorithm generalizes easily. Thus, a suitable value for the learning rate should be low enough so as the network to acquire better precision but high enough that it does not require too much time for the training. Batch size defines the number of images that algorithm examines during each training step. Small batch size indicates that more updates of parameters will take place. But also it bears the risk of lower accuracy, since less images are examined. Again trade-off between the number of updates and the accuracy is needed so as to achieve the best results. Epoch shows the number of times that all the images of the dataset are examined. Number of hidden layers are the layers between the input and the output layer. Many hidden layers can increase the accuracy of the algorithm but also the complexity and the training time, and potentially cause overfitting.

The methods that determine the most appropriate hyper-parameters are the followings.

The first method is the Manual Search which can be applied by observing the impact of the hyper-parameters to results (Goodfellow et al., 2017). Another method is the Random Search. By identifying a range for the hyper-parameters, random selection takes place and then narrowing the range close to those hyper-parameters for the final selection (Bergstra et al., 2012). According to Bergstra et al. (2012), Random Search provides the most accurate results compared to the other methods. The last method is Bayesian Optimisation. This follows the principle of using any existence information for the hyper-parameters from previous experiments to achieve better results (Snoek et al., 2012).

3 Methodology

In this chapter, description of the methods that were used for the training of datasets follows.

3.1 Training from scratch

The first algorithm that was used was for the training of the dataset from scratch. It consisted three convolutional layers, each one followed by a "tanh" activation layer and a max-pooling layer (outlined in Appendix, Chapter 2.1). After, two fully-connected layers were added on the top of it (image 4). All the images were resized to 150x150 and divided into sets of 45%, 30% and 25% for training, validation and testing respectively. Since the dataset was small, data augmentation was used so as to avoid overfitting. Rescaling, shearing, zooming and horizontal flipping were implemented to the dataset. The results that are shown in Tables 3-6 were acquired for 50 epochs and batch size equal to 16. The total time for every training was estimated around 9 hours.

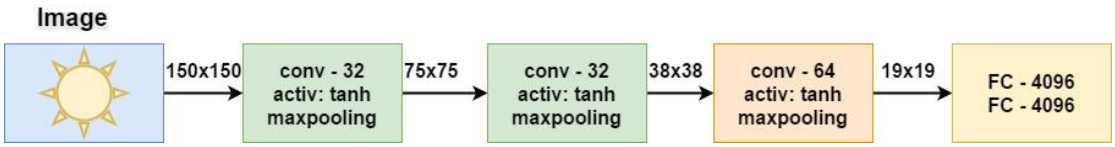

Image 4: The layers that were used for the algorithm from scratch

Another method that was followed for the training of the dataset was Transfer Learning (outlined in Appendix, Chapter 2.2). Different architectures of pretrained networks were used and are described below.

3.2 Inception v3

Inception v3 is a network trained on ImageNet database designed for image recognition tasks (Szegedy et al., 2015). It includes 174 layers which are stacked on top of each other (image 5). Images are placed in two separated folders, melanomas and nevi. Firstly, the pretrained model of Inception v3 is loaded and then bottlenecks (activation maps before the fully-connected layers) for all the images are created. 60% of the images were randomly selected and used exclusively for training, 20% for validation and 20% for the testing. The required input size of images is 299x299. So all images should be resized before the algorithm is applied. During the training, the 2 final fully-connected layers of the network are removed and the new one is added to train the images of the specific project. In every step of the training, according to the batch size, a specific number of images is examined and validation accuracy as well as validation loss are calculated. In the end, this process is repeated once more but only for the test images providing the final results. Even though ImageNet includes around 2 million images, it does not include images of melanomas and nevi but many features are extracted from the other images and assist in the classification of any type of images. The most desirable results came up while learning rate was set up to 0.01 and batch size to 200. The training time was around 45 minutes for each training.

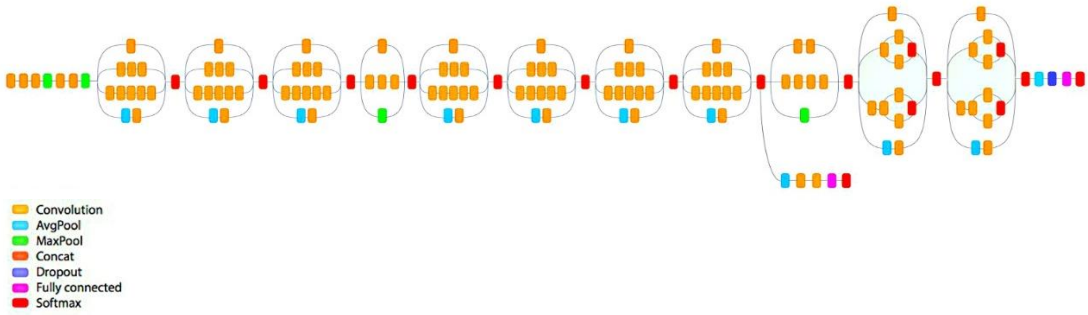

Image 5: The Inception v3 model (Szegedy et al., 2015)

3.3 VGG-16

VGG-16 architecture achieved the second place in ImageNet Large Scale Visual Recognition Competition (ILSVRC) (Russakovsky et al., 2015) of 2014 for the classification task. Its characteristic is that it is consisted of 16 layers using only small (3x3) convolutional filters (Simonyan et al., 2015). Training of this model from scratch can take even weeks because of its large depth. In this project, the VGG-16 architecture was used as a pretrained model and only the fully-connected layers were modified so as to train the specific datasets.

Images were resized to 224x224 for the demands of the model and divided into 45%, 30% and 25% for training, validation and testing respectively. After loading the VGG-16 pretrained model, the last fully-connected layers freeze and the bottlenecks of the new images are created and stored in two numpy arrays. A fully-connected model is connected on top of them and the weights of the new model are saved. Then, the final part of the training takes place in which the method of fine-tuning (outlined in Appendix, Chapter 2.2) is used. The weights from the already trained classifier are used to acquire the final results. Hyper-parameters had the following values: epoch: 50 and batch size: 16. The time for each training was approximately 35 minutes.

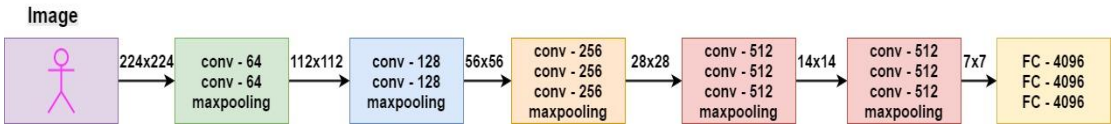

Image 6: The VGG-16 model (Simonyan et al., 2015)

4 Results

4.1 Training from scratch

In the Tables 3 and 4, accuracy, sensitivity and specificity (outlined in Appendix, Chapter 1.2.3) using training from scratch with data augmentation are shown for dermoscopical and clinical images, respectively. The metrics are calculated with regard to melanoma. The algorithm was applied on the datasets that appear in the left column.

Table 3: Results from the training of the datasets of dermoscopical images with data augmentation using the algorithm from scratch with regard to melanoma

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
ISIC Archive (full lesions)	1242	52.37	84.52	69.93
ISIC Archive (all)	1628	83.64	43.37	60.67
ISIC Archive + PH2	1748	57.78	67.50	64.00
Teleskin	837	73.02	49.17	57.38
Teleskin + Menzies	997	57.83	74.00	68.24
Teleskin + ISIC Archive	**2465**	**69.09**	**74.00**	**72.26**
All dermoscopical images	2745	74.55	27.00	43.87

Table 4: Results from the training of the datasets of clinical images with data augmentation using the algorithm from scratch with regard to melanoma

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
Teleskin (clinical)	**697**	**92.96**	**69.01**	**80.99**
Teleskin + Menzies (clinical)	857	60.59	81.94	69.87

Alike the previous Tables (3 and 4), in Tables 5 and 6, the results for accuracy, sensitivity and specificity are presented for dermoscopical and clinical images, without data augmentation this time.

Table 5: Results from the training of the datasets of dermoscopical images without data augmentation using the algorithm from scratch with regard to melanoma

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
ISIC Archive (full lesions)	1242	0	100	60.00
ISIC Archive (all)	1628	0	100	56.02
ISIC Archive + PH2	1748	100	0	35.48
Teleskin	837	0	100	65.57
Teleskin + Menzies	997	100	0	35.62
Teleskin + ISIC Archive	2465	0	100	64.52
All dermoscopical images	2745	0	100	67.75

Table 6: Results from the training of the datasets of clinical images without data augmentation using the algorithm from scratch

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
Teleskin (clinical)	697	0	100	55.37
Teleskin + Menzies (clinical)	857	0	100	56.46

4.2 Inception v3

In Tables 7 and 8, the datasets were trained using transfer learning with the architecture of Inception v3. For every dataset, the mean value of three repetitions is indicated. Also, sensitivity, specificity and accuracy are calculated with regard to melanoma cases.

Table 7: Results from the training of the datasets using Inception v3 architecture for dermoscopical images with regard to melanoma cases

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
ISIC Archive (full lesions)	1242	76.60	78.79	78.1
ISIC Archive (all)	1628	74.68	90.91	83.7
ISIC Archive + PH2	1748	82.18	87.98	84.3
Teleskin	**837**	**87.75**	**92.27**	**91.4**
Teleskin + Menzies	997	87.63	88.82	88.5
Teleskin + ISIC Archive	2465	73.75	89.31	84.1
Teleskin + ISIC Archive + Menzies	2625	83.33	71.46	75.8
All dermoscopical images	2745	82.40	75.40	78.0

Table 8: Results from the training of the datasets using Inception v3 architecture for clinical images with regard to melanoma cases

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
Teleskin (clinical)	**697**	**80.43**	**78.00**	**78.9**
Teleskin + Menzies (clinical)	857	80.51	77.22	78.5

In Table 9, the dataset of Teleskin (837 images: 201 melanomas and 636 nevi), was used so as to show how the different values of hyper-parameters (learning rate and batch size) affect the sensitivity, specificity and accuracy.

Table 9: Results from training of Teleskin dataset for different values of learning rate and batch size with regard to melanoma

Hyper-parameters	Sensitivity(%)	Specificity(%)	Accuracy (%)
learning rate: 0.1, batch size: 200	79.59	95.10	90.23
learning rate: 0.01, batch size: 200	**87.75**	**92.27**	**91.41**
learning rate: 0.001, batch size: 200	88.71	89.37	88.67
learning rate: 0.0001, batch size: 200	89.79	79.22	81.25
learning rate: 0.01, batch size: 10	82.43	93.79	87.15
learning rate: 0.01, batch size: 100	85.36	93.15	89.78
learning rate: 0.01, batch size: 200	**87.75**	**92.27**	**91.41**
learning rate: 0.01, batch size: 500	83.67	92.27	90.63

4.3 VGG-16

In Tables 10 and 11, the results about sensitivity, specificity and accuracy with regard to melanoma for VGG-16 architecture are collected. In Table 10, training took place using dermoscopical images while in Table 11 clinical. In the last one, results from the training for images that were taken from small distance as well as all the clinical images from Teleskin dataset, are presented.

Table 10: Results from the training of the datasets using VGG-16 architecture for dermoscopical images with regard to melanoma

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
ISIC Archive (full lesions)	1242	68.62	75.57	70.33
ISIC Archive (all)	1628	65.77	80.38	73.08
ISIC Archive + PH2	1748	70.93	82.67	77.24
Teleskin	837	57.74	97.17	85.89
Teleskin + Menzies	997	76.25	96.63	90.97
Teleskin + ISIC Archive	**2465**	**78.64**	**95.41**	**89.10**
Teleskin + ISIC Archive + Menzies	2625	57.45	79.29	68.96
All dermoscopical images	2745	59.21	80.00	70.52

Table 11: Results from the training of the datasets using VGG-16 architecture for clinical images, taken from close (< 5 cm) or bigger (5-10 cm) distance with regard to melanoma

Dataset	No of Cases	Sens (%)	Spec (%)	Acc (%)
Teleskin (clinical) (small distance)	476	73.07	88.98	82.65
Teleskin (clinical) (all)	**697**	**84.48**	**88.76**	**86.34**
Teleskin + Menzies (cl) (small distance)	636	86.17	86.73	83.84
Teleskin + Menzies (clinical) (all)	857	78.00	94.18	87.02

5 Discussion

Deep learning algorithms were implemented with the purpose of classifying the given images into melanoma and nevus. In Table 14, results from researches about the use of dermoscopy are presented. Sensitivity and specificity are compared with regard to naked eye examination and dermoscopy. In the last row, results from this work are included, in which results from clinical images are compared to those of naked eye examination and from dermoscopical images to dermoscopy examination. It can be observed that results using deep learning algorithms are better than clinical examination, so those results can be used as an assistant tool for verifying clinicians' diagnosis.

Table 14: Comparison the sensitivity and specificity of this work with clinicians for both naked eye and dermoscopy with regard to melanoma

		Sensitivity (%)		**Specificity (%)**	
Study	**Lesion**	**Naked eye**	**Dermoscopy**	**Naked eye**	**Dermoscopy**
Westerhoff et al. (2000)	Only melanoma	54.6	75.9	-	-
Dolianitis et al. (2005)	Only melanoma	60.9	84.6	85.4	77.7
Argenziano et al. (2006)	All malignant	54.1	79.2	71.3	71.8
Menzies et al. (2009)	All malignant	40.0	55.0	84.6	89.0
Menzies et al. (2009)	Only melanoma	37.5	53.1	84.6	89.0
This work (2017)	**Only melanoma**	**84.5**	**87.8**	**88.8**	**92.3**

The algorithm from scratch was used as a reference point since it is consisted of just few layers. The best results for dermoscopical images were for the dataset of Teleskin combined with ISIC images, providing accuracy 72.26%, sensitivity 69.09% and specificity 74%. For clinical images, the results were quite desirable with accuracy 80.99%, sensitivity 92.96% and specificity 69.01%.

For the algorithm that was built from scratch, images were classified since they underwent data augmentation in one case and without in the other. As it can be seen on Tables 5 and 6, the algorithm without data augmentation could not generalize and classify correctly the images. In all cases, all the images were classified as melanomas or nevi so specificity and sensitivity were 0% respectively.

The method with the best performance for dermoscopical images was proved to be the one using Transfer Learning with the architecture of Inception v3. It achieved accuracy 91.4%, sensitivity 87.75% and specificity 92.27% for the dataset from Teleskin, with learning rate 0.01 and batch size 200. For the clinical images, VGG-16 is considered more appropriate with accuracy 83.84%, sensitivity 86.17% and specificity 86.73%, for the datasets from Teleskin and Menzies with learning rate 0.0001 and batch size 16.

Comparing the results of Inception v3 and VGG-16, it can be observed that for dermoscopical images, the performance of Inception v3 was better than this one of VGG-16, with biggest difference in the values of sensitivity. On the other hand, as far the clinical images, VGG-16 provided better results for all the metrics than Inception v3.

Images from Teleskin and Menzies datasets provided better results compared to those from ISIC Archive and PH2. This can be attributed to the fact that images from Teleskin and Menzies are uniform, non-polarized and in contact to the surface of the skin lesion where oil has been applied. For ISIC Archive and PH2 those details for images are not presented, as a result assumption that images are not uniform, is made.

In Inception v3 model, some trainings took place with use of all images from ISIC Archive and others just with the images of full lesions. The results showed that the training with images of only full lesions had higher sensitivity but lower specificity and accuracy. Also, in trainings that all images from ISIC Archive were used, sensitivity presented low results. This can be attributed to the fact that images are not uniform and skin lesion is not always located in the center of the image.

In Table 9, different values for learning rate and batch size were used so as to compare the results of the algorithm. It is observed that with high learning rate, specificity increases but with trade-off of the sensitivity. This is not so useful since many melanomas are classified as nevi and users can be seriously misguided. For quite small value of learning rate, accuracy and specificity are quite degraded. As far the batch size, very small batch size gives high specificity but low sensitivity. As the batch size increases, the specificity reduces and the sensitivity increases. This happens because more images are taken into consideration on every step of the process and algorithm adopts many features also from images of melanomas. In case of small training batch size, features are extracted mostly from nevi images since they are more in number with result the high specificity.

In Table 11, results for VGG-16 model are presented for clinical images that are divided in two categories, the close ones and the combination of both close and distant images. Results show that there are not big differences but slightly better performance is appeared to be when using all the clinical images. This shows that the algorithm generalizes better when more pictures are available.

In Image 7, there are some examples of misclassified images. The ground truth of the top images is “melanoma” but the label after the implementation of the algorithm was “nevus” and vice versa for the bottom ones. The top left image is a part of a lesion of melanoma appeared to be different than most of the images which are including a full lesion in the center of the image. The top right one appears a full lesion of melanoma in the center of the image, but some hair is included in the image. Bottom left image is a nevus from ISIC Archive dataset which also contains some hair. Bottom right image presents a nevus from Teleskin dataset which probably due to the dark colour got misclassified. So, those factors may lead some images to misclassification.

Image 7: Misclassified images. Top left: part of a lesion of melanoma from ISIC Archive dataset. Top right: melanoma from Teleskin dataset. Bottom left: nevus from ISIC Archive. Bottom right: nevus from Teleskin dataset.

As far the training time of the algorithms, comparing only the different models used Transfer Learning, since the algorithm from scratch demanded much more time, VGG-16 was faster than Inception v3, with not big differences. Probably that happens because it ran in GPU while for Inception v3, Docker was used.

In Table 15, there is a comparison of the best results of this work with other related works. Ramlakhan et al. (2011) used images from cell phone cameras and performed the classification into melanomas by using kNN classifier. In Menegola's et al. (2017) research, VGG-16 model was used and through transfer learning, they classified skin lesions to melanomas and nevi. Some other approaches are described from Yu et al. (2016) and Codela et al. (2016), in which very deep residual networks and ensemble of methods were applied respectively. Also, the results from Nylund's (2016) master's project are presented, which were acquired using an ensemble of methods.

Table 15: Comparison of sensitivity, specificity and accuracy of similar previous approaches with regard to melanoma

Studies	Sensitivity(%)	Specificity(%)	Accuracy(%)
Ramlakhan et al. (2011)	60.7	80.5	66.7
Codela et al. (2016)	62.7	95.7	75.5
Nylund (2016)	77.1	93.0	89.3
Yu et al. (2016)	91.1	95.7	94.9
Menegola et al. (2017)	47.6	88.1	79.2
This work (2017)	**87.8**	**92.3**	**91.4**

Yu et al. (2016) achieved high results for all the metrics. Comparing all the others, this work achieved the highest sensitivity and accuracy. Sensitivity is the most important among metrics since it shows how accurate the prediction of images of melanomas is. Trade-off between sensitivity and specificity took place with purpose of increasing the value of sensitivity. According to Soyer et al. (2007), visual clinical examination provides sensitivity in detecting melanoma between 65% and 80%.

For the deep learning, many images are required so as to achieve more accurate results. In case of medical images, not many are available online in respect of patients' privacy. But by choosing the appropriate architecture of deep learning model as well as suitable hyper-parameters can provide desirable results, as it was shown in this project.

It should be highlighted that the present work is proposed as an auxiliary tool for not experienced physicians so as to strengthen the confidence of their predictions or alternatively for private use to tract the evolution of skin lesions. It cannot replace doctor's examination and clinical visit is required after observing any of the symptoms (outlined in Appendix, Chapter 1.1).

Further research can be focused on classification of different skin lesions broadening the binary classification that was used in this project. Another suggestion is to classify cases of melanomas according to the depth that they are spread. Furthermore, the accuracy of those different stages can be examined and be compared with the others. It would also be interesting, different resolutions to be tested so as to evaluate if accuracy gets affected. Consequently, results of mobile applications for melanoma detection can be more reliable. Finally, bigger collection of images can administer results with higher accuracy.

6 Conclusion

The aim of the project was to classify, with the use of Deep Learning, images of melanomas and nevi. The results of the project using the pretrained models of Inception v3 and VGG-16 for both dermoscopical and clinical images were found to be comparable and even better than those of dermatologists. So, an application using the present algorithms can provide reliable results and serve as an assisting tool from dermatologists and self-examination tool for first diagnosis from individual users.

Appendix - State of the Art

A.1 Melanoma

Over the last 30 years, skin cancer has appeared to be more frequent than all other types of cancer combined (Stern R., 2010). Skin cancer is divided into melanocytic and non-melanocytic form. Melanocytic lesions include melanoma (malignant) and melanocytic nevus (benign) and non-melanocytic lesions include basal cell carcinoma, squamous cell carcinoma (malignant) and seborrheic keratosis, vascular lesions and dermatofibroma (benign).

Melanoma is a disease in which malignant cells form in melanocytes. Melanocytes are the cells that produce melanin, which gives the colour to the skin. They make more pigment when the skin is exposed to sun or artificial light which causes the skin to darken (NCI's PDQ cancer information summary, 2016). The most common types of melanoma are superficial spreading melanoma, nodular melanoma, acral lentiginous melanoma, and lentigo maligna melanoma. (Tsao H. et al., 2012).

According to Skin Cancer Foundation (2017), 87,110 new cases of invasive melanoma are estimated to be diagnosed only in the U.S in 2017 and around 9,730 people to die in U.S. of melanoma in 2017. In Table 1, the number of incidences of melanoma, mortality and 5-year prevalence for some countries of Europe and Australia are presented. The main risk factors for melanoma are long periods of exposure to natural or artificial sunlight (such as tanning beds), having a fair complexion, having family or personal history of melanoma, having a weakened immune system and having several large or many small moles (NCI's PDQ cancer information summary, 2016).

Table 1: Melanoma incidence, mortality and 5-year prevalence in Europe (Ferlay et al., 2012 & Bray et al.,2012) and Australia (Australian Institute of Health and Welfare, 2015 & Australian Bureau of Statistics, 2015)

Country	**Incidence**		**Mortality**		**5 – year prevalence**
	Number	**Rate per 100000**	**Number**	**Rate per 100000**	
EU (27 countries) (Ferlay et al., 2012 & Bray et al.,2012)	82075	13,0	22000	2,2	323467
Denmark	1596	24,1	228	3,0	6443
Germany	16884	14,8	2671	2,0	66997
UK	14445	19,0	2195	2,6	57163
Norway	1506	25,3	325	3,1	6247
Sweden	2911	23,9	565	4,0	12088
Switzerland	2484	25,8	384	3,5	10357
Netherlands	4804	24,4	853	3,9	20223
Australia (Australian Institute of Health and Welfare, 2015 & Australian Bureau of Statistics, 2015)	12960	49,0	1617	6,5	48364

A.1.1 Symptoms

Normal moles usually have appeared during first years of life and are even in colour and quite small. New spots on the skin or a spot which is changing in colour, size or shape are the most important warning signs of melanoma. Also, spots that look different than other spots on the skin, known as the ugly duckling sign, is another warning sign of melanoma.

Furthermore, melanoma possibly will display those symptoms: a sore that doesn't heal, spread of pigment from the border of a spot into surrounding skin, redness or a new swelling beyond the border of the mole, change in sensation, such as itchiness, tenderness, or pain and change in the surface of a mole – scaliness, oozing, bleeding, or the appearance of a lump or bump (Brandberg, 1992).

Finally, some other warning signs that could also present melanoma are a dark stripe under a toe or fingernail, a slow developing part of the skin that resembles a scar or an area of dark skin around a toenail or fingernail.

A.1.2 Diagnosis

Although melanoma is a rare form of skin cancer, its fatality rate is high and it is more possible to spread to nearby tissues and other parts of body compared to other types of skin cancer. According to Skin Cancer Foundation (2017), it accounts for less than one percent of skin cancer cases, but for the majority of deaths caused by skin cancer. The estimated 5-year survival rate in cases of early detection of melanoma is around 98% in U.S. Same rate falls to 62% when the disease reaches the lymph nodes and 18% when it metastasizes to distant organs (Skin Cancer Foundation, 2017). So, diagnosis in early stage is of vital importance for the treatment and cure of melanoma. The most common types of examinations are clinical and dermoscopical, as well as automated procedures, that have been rapidly developed over the last years. Specific characteristics that are taken into consideration in clinical and dermoscopical examinations so as to detect melanoma, are described in next sections. More detailed analysis about automated procedures is found in "Deep learning" chapter.

A.1.2.1 Clinical examination

Clinical examination includes complete cutaneous examination which should be performed on every patient regardless the age, since dysplastic nevi (moles with appearance different than common moles) and melanomas may appear on any area of the skin (Soyer H.P. et al., 2007). Examination every 6 months is required for patients with history of dysplastic nevi, non-melanoma skin cancer or melanoma. The patient should lie in a horizontal position on the examining table, and be examined in both supine (lying on back) and then prone (lying on chest) positions. Examination of the scalp takes place with the use of a blowing device for better viewing. For the complete cutaneous examination, also examination of intertriginous areas is required, including the axilla, groin, interdigital webs of hands and feet as well as nail apparatus.

Diagnosis of melanoma is based on the following methods in which specific features of skin lesions are taken into account (Soyer H.P. et al., 2007).

ABCDE criteria is a simple mnemonic acronym that alerts the attention on some of the key features of melanoma, Asymmetry, Border irregularity, Colour, Diameter and Evolving.
Three Cs method is also used for melanoma detection which is standing for Colour, Contour and Change.
The Glasgow 7-point checklist diagnostic method includes: (a) change in size; (b) irregular shape; (c) irregular colour (major criterion); (d) diameter at least 7 mm; (e) inflammation; (f) oozing/bleeding; and (g) change in sensation (minor criterion).

Furthermore, total cutaneous photography, especially in case of patients with many melanocytic nevi, is essential which is helpful in identifying new melanocytic lesions as well as significant changes in pre-existing ones.

Clinical examination is based on naked eye examination so several details may not be observed. The accuracy of clinical examination is only about 60% (Kittler. H. et al., 2002). To reduce those limitations and improve the accuracy, dermoscopical examination can be performed which is analysed in the next section.

A.1.2.2 Dermoscopical examination

Dermoscopy, also known as dermatoscopy, epiluminescence microscopy and amplified surface microscopy, is an in-vivo method that is used for recognition of melanoma or differential diagnosis of pigmented lesions of the skin (Soyer H.P. et al., 2007). Handheld devices (dermoscopes) with 10-20 times magnification are used to visualise the surface of the skin lesion. Usually, a glass plate is placed on the surface that will be examined and provides an even surface. Immersion liquids, such as alcohol or cosmetic gel, are applied for better image quality. Alternatively, devices that use polarized light are used so as to reduce the surface reflections.

After the identification of melanocytic lesion, the decision has to be made if the lesion is benign or malignant. This can be accomplished with the use of the following algorithms (Soyer H.P. et al., 2007).

Pattern analysis

Pattern analysis has been used by clinicians and dermatologists to distinguish among different patterns and differentiate the case to benign or malignant. Some of the most common characteristics are atypical pigment network, streaks (radial streaming and pseudopods), negative pigment network, crystalline structures, atypical dots, and globules, off centered blotches, regression structures, blue white veil on raised areas, atypical vascular structures and peripheral brown structureless areas (Serrano et al., 2009). The combination of three or more patterns is called multicomponent pattern and it is usually sign of melanoma. In figures 1 and 2, patterns for benign nevi and melanoma respectively are presented.

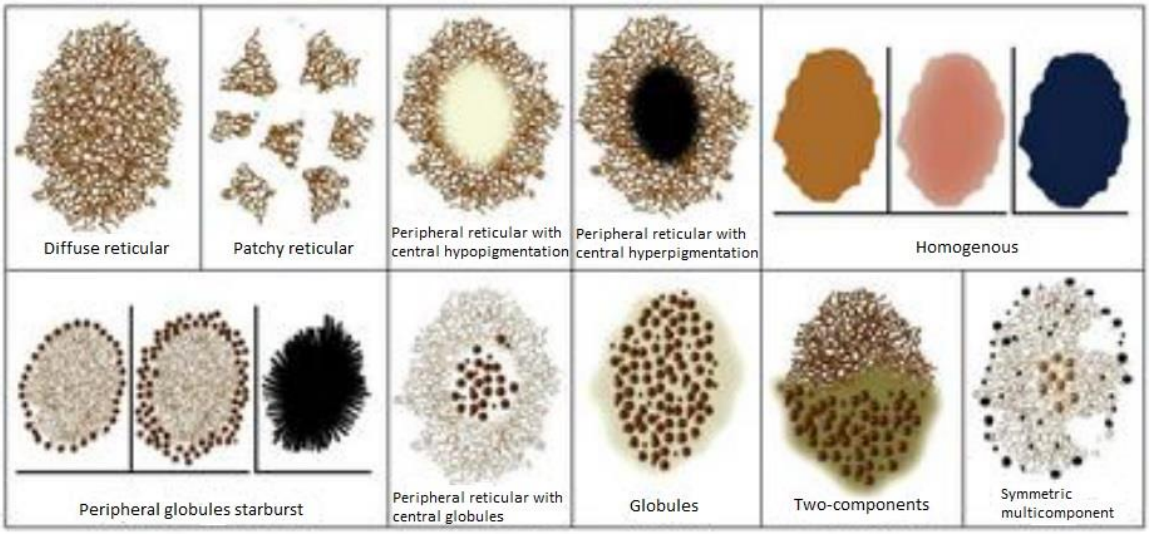

Figure 1: Diagram showing Dermoscopy benign nevus patterns (Marghoob A. & Braun R P., 2012)

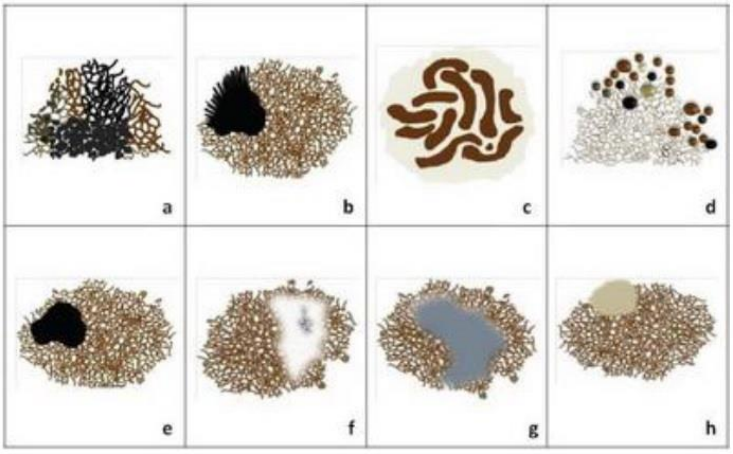

Figure 2: Diagram showing melanoma specific features (Marghoob A., 2012)

ABCD Rule of Dermoscopy

Like in clinical examination, ABCD rule of dermoscopy is used for the recognition of melanomas (Nachbar et al., 1994). It is a scoring system by giving points to parameters for asymmetry, border, color and dermoscopic structures like it is shown in Table 2. After, by multiplying the points with the relevant weight factor, a sub-score for each parameter is calculated and finally the sum of them equals to a total score.

Table 2: ABCD Rule of dermoscopy (Soyer H.P. et al., 2007)

		Points	Weight factor	Sub-score range
Asymmetry	Complete symmetry	0	1.3	0−2.6
	Asymmetry in one axis	1		
	Asymmetry in two axes	2		
Border	Eight segments, one point for abrupt cut-off of pigment	0-8	0.1	0-0.8
Color	One point for each color: white; red; light brown; dark brown; black; blue gray	1−6	0.5	0.5−3.0
Dermoscopic structures	One point for every structure: pigment network; structureless areas; dots; globules; branched streaks	1-5	0.5	0.5-2.5
		Total score range:		1.0–8.9

A lesion with total score lower than 4.75 can be considered benign. Higher than 5.45 is considered malignant and in this case the lesion should be removed. For total score between 4.75 and 5.45 the lesion is considered suspicious and its evolution should be examined.

7-point checklist

Another algorithm is the 7-point checklist. Three characteristics are considered as major (atypical pigment network, blue-whitish veil, atypical vascular pattern) with score 2 points each and four characteristics are of minor importance and their score is 1 point. Total score equal or

higher than 3 indicates melanoma. In Table 3, the characteristics with relevant scores are presented.

Table 3: 7-point checklist method (Soyer H.P. et al., 2007)

Criteria	7-point score
Major	
Atypical pigment network	2
Blue-whitish veil	2
Atypical vascular pattern	2
Minor	
Irregular streaks	1
Irregular pigmentation	1
Irregular dots/globules	1
Regression structures	1

Menzies method

Menzies method for diagnosing melanoma is based on the absence of both following negative features: single colour (tan, dark brown, gray, black, blue, and red, but white is not considered) as well as point and axial symmetry of pigmentation (Argenziano et al., 2001). Additionally, a lesion can be considered as melanoma if at least one of the positive features from table 4 is observed.

Table 4: Menzies method (Soyer H.P. et al., 2007)

Negative features
Point and axial symmetry of pigmentation
Presence of a single color
Positive features
Blue-white veil
Multiple brown dots
Pseudopods
Radial streaming
Scar-like depigmentation
Peripheral black dots−globules
Multiple colors (five or six)
Multiple blue/gray dots
Broadened network

3-point checklist

In three-point checklist only three dermoscopic criteria are examined so as to investigate any suspicious malignancy. Those are i. asymmetry, ii. atypical network and iii. blue-white structures. When two or three features are present, the lesion can be identified as malignant. With this method, it does not occur differentiation between melanocytic and non-melanocytic lesions but it is a screening technique for recognition of malignancies, including melanoma and basal cell carcinoma. This algorithm is not recommended by IDS (International Dermoscopy Society).

A.1.2.3 Evaluation metrics

To be able to compare the previous methods and evaluate their performance, operationalization is needed. Except for the accuracy that shows how many cases have correctly been diagnosed, sensitivity and specificity are evaluation metrics that can give a more complete representation of evaluation of clinical tests. The possible results after a diagnosis compared to ground truth can be the following:

Ground Truth\ Classified as	**Positive**	**Negative**
True	True Positive	True Negative
False	False Positive	False Negative

Sensitivity and specificity are calculated by the following formulas (Parikh R. et al., 2008):

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

So, sensitivity shows the proportion of cases with this condition that have been correctly identified, taking into consideration the total number of cases with the specific condition. Accordingly, specificity represents the proportion of cases without the condition that have been correctly identified, taking into consideration the total number of cases without the specific condition.

Comparison of dermoscopical methods

Dolianitis C. et al. (2005) performed an experiment with the aim to evaluate the accuracy of dermoscopical methods. The participants were medical practitioners who were trained in Pattern analysis, ABCD, 7-point checklist and Menzies method. In Table 5 the results of the experiment are presented, comparing sensitivity, specificity and accuracy. The accuracy and sensitivity are higher in Menzies Method while higher specificity is acquired by Pattern analysis, as can be seen in Table 5. Both sensitivity and specificity are important metrics since a rate for correctly identified those with disease and those without respectively is provided. So, for better results for melanoma recognition both methods should be used.

Table 5: Comparison of dermoscopical algorithms (Dolianitis C. et al., 2005)

Method	**Sensitivity**	**Specificity**	**Accuracy**
Pattern analysis	68,4%	85,3%	76,8%
ABCD	77,5%	80,4%	79,0%
7-point checklist	81,4%	73,0%	77,2%
Menzies method	84,6%	77,7%	81,1%

Apart from the medical part for diagnosis, prognosis and treatment of melanoma recent years many attempts have taken place for development and improvement of automated melanoma detection methods. Those methods provide the advantage to public to have the ability for self-examination as well as an aid tool for clinicians in decision making. It should be noted that those algorithms cannot replace clinicians and their purpose is the early detection of possible skin cancer. In next section, main features of deep learning are presented which are used in building algorithms for melanoma detection.

A.2 Deep learning

Learning is the ability of acquiring knowledge through being taught, study or experience. The field of machine learning refers to the computational processes that control learning in both humans and machines. Machine learning can deal with problems that require intelligence, including tasks related to diagnosis, motor control, planning and natural language (Langley P., 1996).

Deep learning is a new field of Machine Learning which is based on algorithms with objective of deriving features from the sensory signals and help to make sense of data, such as images, text and speech (Deng L. and Yu D., 2013). Deep learning architectures are composed of multiple levels with several hidden layers. Its aim is through algorithms to indicate to machine how to change its parameters that are used in each layer from the ones in the previous layer. In a simple case, a set of neurons receives an input signal and passing through one layer, another set produces the output signal (fig 3, left). In reality, usually there are more than one hidden layers and every output is used as input for the next layer (fig 3, right). Many hidden layers offer the advantage of solving complex pattern recognition problems but often they are harder to train (Nielsen M., 2015). So, according to the problem, different number of hidden layers is needed.

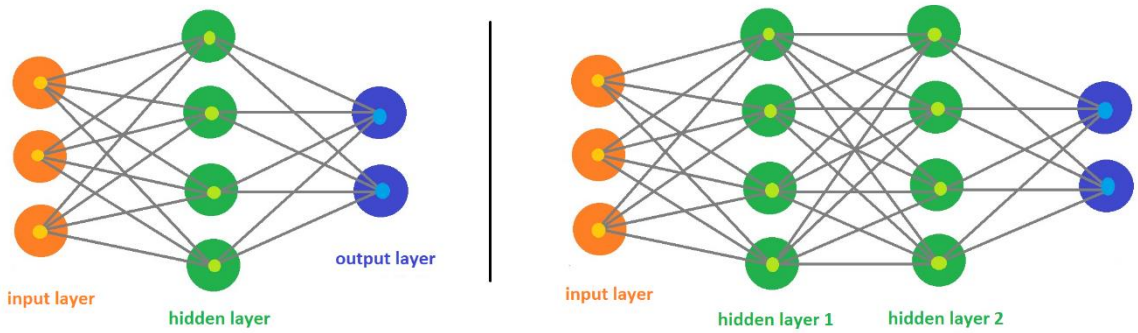

Figure 3: Neural Networks with 1 and 2 hidden layers respectively

A.2.1 Layers

As described above, Neural Networks are consisted of layers which transform the input signal to another through several functions. The three main types of layers of a typical neural network are Convolutional Layer, Pooling Layer and Fully-Connected Layer. (Cs231n.github.io, 2017)

A.2.1.1 Convolutional layer

Convolutional layer is responsible for the most computational heavy lifting. The parameters of this layer are consisted of a set of learnable filters. Every filter is small spatially, for example a typical filter on a first layer might have size 5x5, but it extends through the full depth of the input volume, which for images is 3 that represents the colour channels, so 5x5x3. Then, each filter is slid across the input volume, and dot products between this volume and the filter are computed at any position, producing a 2-dimensional activation map (Figure 4). In each convolutional layer, every filter will produce a unique activation map and finally all of them will be combined to produce the output volume.

So, in other words, in convolutional layer the computation of the output of neurons by computing a dot product between their weights and a small region of the input volume, takes place. The filters are convolved with the image producing activation maps at every spatial position which leads to the output volume. This is why the layer is called convolutional layer.

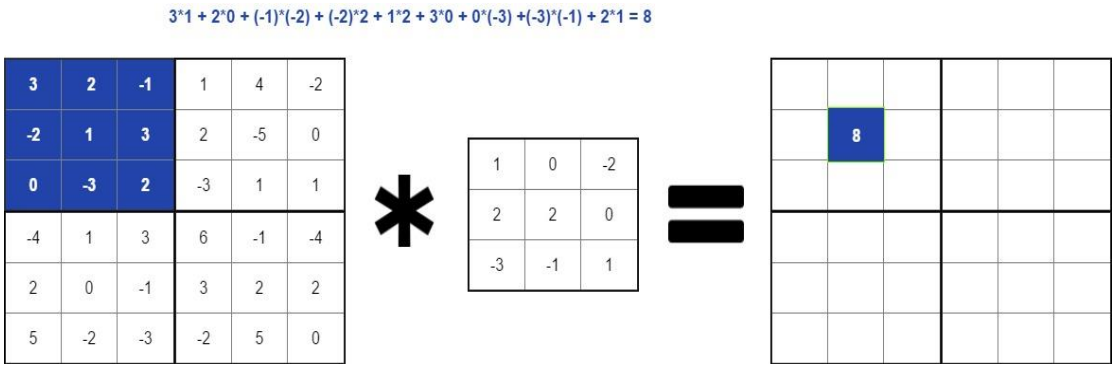

Figure 4: How convolutional layer works

A.2.1.2 Pooling layer

Pooling layer performs a downsampling operation along weight and height resulting in smaller representations and thus, becoming more manageable (Figure 5). Pooling layers are usually between convolutional layers with the aim to reduce the amount of parameters and computations and therefore to control overfitting (common problem in machine learning, when the algorithm predicts well in training but fails in testing). The operation takes place in each activation map independently and by using the MAX function, it resizes it spatially.

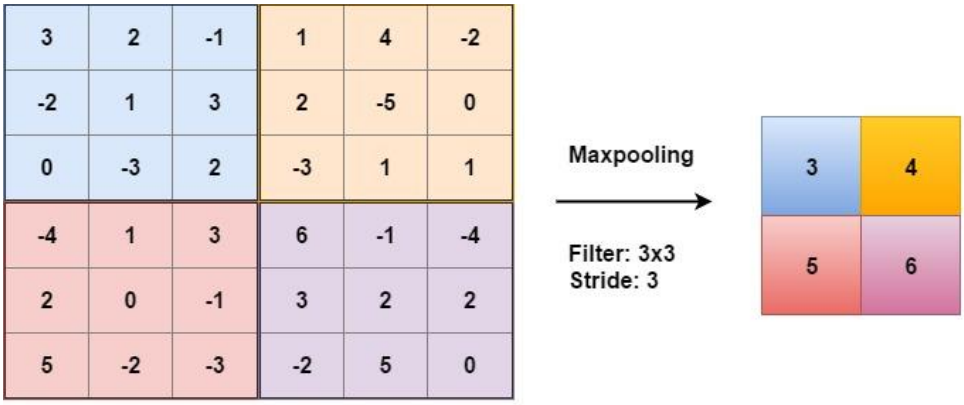

Figure 5: How max-pooling layer works

A.2.1.3 Fully Connected Layer (FC layer)

In a Fully Connected layer, neurons have full connections to all activations in the previous volume. So, those activations can be computed as the result of a matrix multiplication followed by a bias offset. (Figure 6)

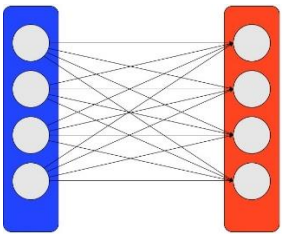

Figure 6: Fully connected layer

A.2.1.4 Softmax

In case that more than two classes have to be classified, Softmax is used as the last layer. The Softmax function is given by (Nielsen M., 2015):

$$a_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

Where a_j is the activation of the j_{th} output volume and z the weighted input volumes. So, for example for input [1, 2, 3, 4, 1, 2, 3], the softmax values according to the previous equation will be [0.024, 0.064, 0.175, 0.475, 0.024, 0.064, 0.175], with greater value the one that is related to 4 (the more weighted one). The sum of the values of Softmax always equals to 1. Finally, the greater value will be used to show in which class the input volume belongs.

A.2.2 Transfer learning

In most problems, the size of the data is not big enough and a lot of time is required to train a convolutional network from the scratch. For this reason, it is common to use a pretrained network on a very large dataset (such as ImageNet (Szegedy et al., 2014) in 1.2 million images) and then use it either as an initialization or a fixed feature extractor for the task of interest (Weiss et al., 2016). This method is called Transfer learning. For example, from a network trained to classify types of trees, by removing the last layer of the algorithm and retrain it with new layer for types of flowers, the classification of the flowers can be achieved in much less time than of trees. The most common ways to apply transfer learning are the following (Cs231n.github.io., 2017):

Fixed Feature extractor

On a pretrained network, remove the last fully-connected (FC) layer and use the rest as a fixed feature extractor for the new dataset. In the end, train a linear classifier (linear Support Vector Machine or Softmax classifier). As it can be seen in Figure 4, the last FC layer which is in green box will be replaced by the new one and softmax will be retrained. This method is appropriate especially for small datasets.

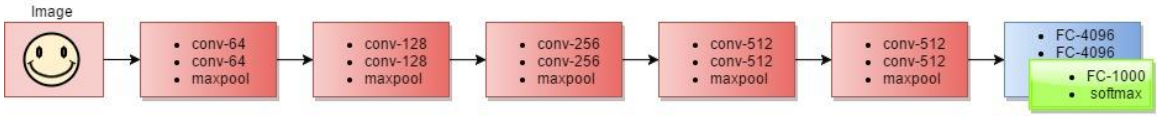

Fig 4: Fixed Feature extractor

Fine-tuning

The other method relies on replacing and retraining the classifier for the new dataset and then to fine-tune the weights of the pretrained network by continuing the backpropagation. In Figure 5, green box includes the layers that should be replaced for fine-tuning. There are the options of fine-tuning all the layers or just some of the higher-levels of the network so as to avoid overfitting. This is helpful because the earlier layers contain more generic features than the later ones that are more specific and detailed. It is appropriate for medium to large sized datasets.

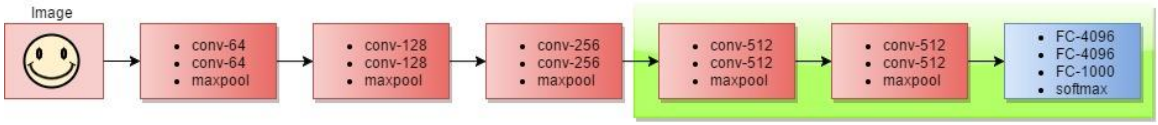

Fig 5: Fine-tuning

A.2.3 CPU and GPU

Central Processing Unit (CPU) is the main part for operations in computer and Graphics Processing Unit (GPU) is traditionally used mostly for the acceleration of creation of images. Lately, because of the disability of CPUs to perform very complex computations, further capacities of GPUs have been explored.

The design of GPUs offers the ability of carrying out huge computations which are performed on the data level parallelism. The advantages of GPUs are that they are fast and cheap as well as energy efficient. On the other hand, it is hard to program them, there are no mature industrial standard models and not all algorithms can be executed faster than in CPUs.

Furthermore, CPUs are more appropriate at handling single calculations quickly since they have a faster clock speed while GPUs are more efficient at handling multiple calculations because of the number of cores they include. So, applications that can be performed well in GPUs are 3D graphics processing based on rasterization and physics simulations and some that cannot perform easily on GPUs are ray tracing and compiling, in particular scanning and parsing.

As a conclusion, CPUs and GPUs have several differences but also many similarities. The combination of them can produce the best results in most demanding problems but the transition from CPU to GPU for everyday tasks is quite difficult since different programming techniques are required.

A.2.4 Theano and Tensorflow

Theano is a library written in Python that allows the definition, optimization and evaluation of mathematical expressions involving multi-dimensional arrays and can take advantage of using GPU to improve the speed (Bastien et al., 2012). When it is used with other libraries, it can be well suited for research and data exploration (Bergstra et al., 2010).

Tensorflow is also a Python open source software library, created from Google to replace Theano for numerical computations using data flow graphs (Abadi et al., 2016). It also provides the ability to deploy computations to one or more CPUs or GPUs with a single API (Abadi, 2016). The system has a wide variety of applicability, such as by Airbnb, Google, Twitter, Snapchat and Dropbox.

Theano and Tensorflow are quite similar libraries, using Python and Numpy and nice computational graph abstraction. Theano also uses high level wrappers (Keras, Lasagne) that facilitate the processes. On the other hand, Tensorflow provides the tool "Tensorboard" which is very useful for visualization, is able for data and model parallelism and compiles faster than Theano. Main disadvantages of Theano are that error messages often can be unhelpful, can have long compiling time for large models and it can use only one GPU. For Tensorflow, main disadvantages include not many pretrained models, computational graph is pure Python, so it is slow and dynamic typing is error-prone on large software projects (Bergstra et al., 2010).

Both libraries have their advantages and disadvantages. So the choice of library depends on the demands of the task and the ability on CPU and GPU of the available computer.

A.2.5 Keras

Keras is a high-level neural networks library, written in Python and capable of running on top of either TensorFlow or Theano (Keras, 2017). It was developed with the purpose of enabling fast experimentation. Providing results with the least possible delay, Keras constitutes a key for good research. It is designed on the following properties: modularity, so many functions, graphs and neural layers to create new models when combined together, minimalism, that it is important each module to be short and simple and easy extensibility, the ability to add new models.

References

Abadi, M. (2016). TensorFlow: learning functions at scale. ACM SIGPLAN Notices, 51(9), pp.1-1.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv: 1603.04467v2.

Argenziano G, Puig S, Zalaudek I, Sera F, Corona R, Alsina M, et al. (2006). Dermoscopy improves accuracy of primary care physicians to triage lesions suggestive of skin cancer. J Clin Oncol. 2006;24(12):1877–82.

Argenziano, G. and Soyer, H. (2001). Dermoscopy of pigmented skin lesions – a valuable tool for early. The Lancet Oncology, 2(7), pp.443-449.

Atlasdermatologico.com.br. (2017). Dermatology Atlas. [Online] Available at: http://www.atlasdermatologico.com.br/ [Accessed 31 Mar. 2017].

Australian Bureau of Statistics (2013). 3303.0 - Causes of Death, Australia. Canberra: Australian Bureau of Statistics.

Australian Institute of Health and Welfare (2015). Australian Cancer Incidence and Mortality (ACIM) books: Melanoma of the skin. Canberra: AIHW.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D. and Bengio, Y. (2012). Theano: new features and speed improvements.

Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures. arXiv: 1206.5533v2

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research. Montreal, Canada.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D. and Bengio, Y. (2010). Theano: A CPU and GPU Math Compiler in Python.

Brandberg, Y., Bolund, C., Sigurdardottir, V., Sjödén, P. and Sullivan, M. (1992). Anxiety and depressive symptoms at different stages of malignant melanoma. Psycho-Oncology, 1(2), pp.71-78.

Bray F, Ren JS, Masuyer E, Ferlay J. (2013). Estimates of global cancer prevalence for 27 sites in the adult population in 2008. Int J Cancer. 1;132(5):1133-45. doi: 10.1002/ijc.27711. Epub 2012 Jul 26.

Codella, N., Nguyen, Q., Pankanti, S., Gutman, D., Helba, B., Halpern, A. and Smith, J. (2016). Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. Ibm Journal of Research and Development.

Cs231n.github.io. (2017). CS231n Convolutional Neural Networks for Visual Recognition. [Online] Available at: http://cs231n.github.io/ [Accessed 1 Mar. 2017].

Deng L. and Yu D. (2013). Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing, vol. 7, nos. 3–4, pp. 197–387.

Dermquest.com. (2017). DermQuest: Home. [Online] Available at: https://www.dermquest.com/ [Accessed 31 Mar. 2017].

Dolianitis, C., Kelly, J., Wolfe, R. and Simpson, P. (2005). Comparative Performance of 4 Dermoscopic Algorithms by Nonexperts for the Diagnosis of Melanocytic Lesions. Archives of Dermatology, 141(8).

Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, Forman D, Bray F. (2013). Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 49(6):1374-403. doi: 10.1016/j.ejca.2012.12.027.

Goodfellow, I., Bengio, Y. and Courville, A. (2017). Deep learning. 1st ed. Cambridge, Mass: The MIT Press, pp.428-432.

Gorantla, V. and Kirkwood, J. (2014). State of Melanoma. Hematology/Oncology Clinics of North America, 28(3), pp.415-435.

Isic-archive.com. (2017). ISIC Archive. [Online] Available at: https://isic-archive.com/ [Accessed 31 Mar. 2017].

Keras (2017). Keras Documentation. [Online] Available at: https://keras.io/ [Accessed 1 Mar. 2017].

Kittler, H., Pehamberger, H., Wolff, K. and Binder, M. (2002). Diagnostic accuracy of dermoscopy. Austria: The Lancet Oncology, 3(3), pp.159-165.

Langley, P. (1996). Elements of machine learning. 1st ed. San Francisco, Calif.: Morgan Kaufmann.

Marghoob A. (2012) Pattern analysis. Atlas of Dermoscopy. London: Informa Healthcare.

Marghoob A, Braun R P. (2012) Two-step algorithm: Differentiating melanocytic from nonmelanocytic lesions. Atlas of Dermoscopy. London: Informa Healthcare.

Mendonça T., Ferreira P. M., Marques J., Marcal A. R. S., Rozeira J. (2013). PH² - A dermoscopic image database for research and benchmarking, 35th International Conference of the IEEE Engineering in Medicine and Biology Society, July 3-7, 2013, Osaka, Japan.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F., Avila, S. and Valle, E. (2017). Knowledge Transfer for Melanoma Screening with Deep Learning. Arxiv: 1703.07479

Menzies SW, Emery J, Staples M, Davies S, McAvoy B, Fletcher J, et al. (2009). Impact of dermoscopy and short-term sequential digital dermoscopy imaging for the management of pigmented lesions in primary care: a sequential intervention trial. Br J Dermatol. 2009;161(6):1270–7. Epub 2009 Jun 27.

Menzies, S. (2009). Dermoscopy. 3rd ed. Sydney: McGraw-Hill.

Miller, K. D., Siegel, R. L., Lin, C. C., Mariotto, A. B., Kramer, J. L., Rowland, J. H., Stein, K. D., Alteri, R. and Jemal, A. (2016), Cancer treatment and survivorship statistics, 2016. CA: A Cancer Journal for Clinicians, 66: 271–289. doi:10.3322/caac.21349

Nachbar F1, Stolz W, Merkle T, Cognetta AB, Vogt T, Landthaler M, Bilek P, Braun-Falco O, Plewig G. (1994). The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions. Journal of the American Academy of Dermatology.

Nielsen A. Michael (2015). "Neural Networks and Deep Learning", Determination Press.

Nylund A. (2016). To be, or not to be Melanoma. Convolutional neural networks in skin lesion classification. Degree project in medical engineering. Stockholm, Sweden.

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology, 56(1), p.45.

Parkin, D. M., Pisani, P. and Ferlay, J. (1999), Global cancer statistics. CA: A Cancer Journal for Clinicians, 49: 33–64. doi:10.3322/canjclin.49.1.33

Ramlakhan, K. and Shang, Y. (2011). A Mobile Automated Skin Lesion Classification System. 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), pp.211-252.

Serrano, C. and Acha, B. (2009). Pattern analysis of dermoscopic images based on Markov random fields. Pattern Recognition, 42(6), pp.1052-1057.

Siegel, R., Ma, J., Zou, Z. and Jemal, A. (2015). Cancer statistics, 2015. CA: A Cancer Journal for Clinicians, 64(1), pp.9-29.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556.

Skin Cancer Foundation (2017). Skin Cancer Facts & Statistics - SkinCancer.org. [Online] Available at: http://www.skincancer.org/skin-cancer-information/skin-cancer-facts [Accessed 25 Feb. 2017].

Snoek J., Larochelle H., and Adams R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems 25: 2960-2968

Soyer, H., Argenziano, G., Hofmann-Wellenhof, R. and Johr, R. (2007). Color Atlas of Melanocytic Lesions of the Skin. 1st ed. Berlin: Deutsches MAB-Nationalkomitee beim Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit.

Stern, R. (2010). Prevalence of a History of Skin Cancer in 2007: Results of an Incidence-Based Model. Archives of Dermatology, 146(3).

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. (2015). Rethinking the Inception Architecture for Computer Vision. arXiv: 1512.00567v3

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2014). Going Deeper with Convolutions. arXiv: 1409.4842.

Tsao, H., Chin, L., Garraway, L. and Fisher, D. (2012). Melanoma: from mutations to medicine. Genes & Development, 26(11), pp.1131-1155.

Weiss, K., Khoshgoftaar, T. and Wang, D. (2016). A survey of transfer learning. Journal of Big Data, 3(1).

Westerhoff K, McCarthy WH, Menzies SW. (2000) Increase in the sensitivity for melanoma diagnosis by primary care physicians using skin surface microscopy. Br J Dermatol. 2000;143(5):1016–20

Yu, L., Chen, H., Dou, Q., Qin, J. and Heng, P. (2016). Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. IEEE Transactions on Medical Imaging, 36(4), pp.994-1004.

TRITA 2017:74