Deep Learning in Melanoma Screening

Aprendizado Profundo em Triagem de Melanoma

Campinas
2018
Afonso Menegola

Deep Learning in
Melanoma Screening
Aprendizado Profundo em
Triagem de Melanoma

Master’s dissertation presented to the Graduate Program of the School of Electrical and Computer Engineering of the University of Campinas to obtain a Master’s degree in Electrical Engineering, in the area of concentration of Computer Engineering.

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na área de concentração de Engenharia de Computação.

Supervisor: Prof. Dr. Eduardo Alves do Valle Junior
Co-supervisor: Dra. Lin Tzy Li

Este exemplar corresponde à versão final da tese defendida pelo aluno Afonso Menegola, e orientada pelo Prof. Dr. Eduardo Alves do Valle Junior

Campinas
2018
Menegola, Afonso, 1991-  
M524d  
Deep learning in melanoma screening / Afonso Menegola. – Campinas, SP : [s.n.], 2018.

Orientador: Eduardo Alves do Valle Junior.  
Coorientador: Lin Tzy Li.  
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.


Informações para Biblioteca Digital

Título em outro idioma: Aprendizado profundo em triagem de melanoma  
Palavras-chave em inglês:  
Neural networks  
Machine learning  
Melanoma  
Image  
Área de concentração: Engenharia de Computação  
Titulação: Mestre em Engenharia Elétrica  
Banca examinadora:  
Eduardo Alves do Valle Junior [Orientador]  
Leticia Rittner  
Cristina Nader Vasconcelos  
Data de defesa: 23-02-2018  
Programa de Pós-Graduação: Engenharia Elétrica
Candidato: Afonso Menegola RA: 163686
Data da Defesa: 23 de fevereiro de 2018

Título da Tese: Deep Learning in Melanoma Screening (Aprendizado Profundo em Triagem de Melanoma).

Prof. Dr. Eduardo Alves do Valle Junior (Presidente, FEEC/UNICAMP)
Profa. Dra. Leticia Rittner (FEEC/UNICAMP)
Profa. Dra. Cristina Nader Vasconcelos (IC/UFF)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no processo de vida acadêmica do aluno.
Acknowledgements

I would like to thank my supervisor, Eduardo Valle, who is an outlier. He has confidence of achieving greatness, and he is not afraid of get his hands dirty on experiments to do it. Because of this, we made partnership in the experiments that led our group to the top score on the melanoma classification competition. I find it really valuable someone who not only open up our eyes for the paths of science, but who also proactively follow this path together. I would also like to thank him for strongly emphasizing us to NOT contaminate train and test sets.

Lin Tzy Li, my co-supervisor, is an example to follow. She has all the professionalism traits I want to have when I grow up. Always writing down what is happening in our group meetings (which made me bought my own notebook to start writing my own notes), proactively participating in text reviews, suggestions, wanting the world to be changed and seeing our group as a way to do it.

Michel Fornaciali played one of the most important roles on this work: the facilitator. As I was running the experiments, there he was in the battle’s front line, documenting the articles, performing reviews, helping with random problems that ALWAYS appears. And all of this with a lot of fun, trading memes on our Slack’s private chat, laughing about our situation. It was really a pleasure to work with you, Michel.

From micro to macro, I would like to thank all Titans (research group led be Eduardo) and RECOD (lab from which I belong, offering all computational resources) team members, specially Sandra Avila, who is one of the the best professors UNICAMP could possibly have: creative, professional, student oriented, a true researcher. Never lose your touch, Sandra. I also thank Ramon Pires, for all the help in the first semester, explaining and sharing his code that I used to run the experiments on the conference paper.

It is really difficult for me to decide who from my family I would thank first. On one hand, I have my parents, João Menegola Sobrinho e Laurinda Menegola, who I don’t even have words to describe, they are the people I love the most, who made me who I am, who always invested on me, believed on me. They are my safe haven. And their food, oh my god, their food. If you want to feel their love, you need to eat their food, it’s like a direct streamline from your stomach to your heart. On the other hand, I have my partner Priscila Bassani, my point of support for everything in my life. She is the one who I share the happy and sad moments, and no matter which one, she is always there, cheering and being patient. Just like I wrote on my undergraduate thesis, thank you for your limitless support. Last but not least, I thank my awesome brothers, Júlia and Bruno Menegola,
who always were and always will be my inspirations for my adult life.

During this two years, I met two people who helped me to shape the Afonso who is starting a career in the industry. Ramon Oliveira and Pedro Tabacof were the ones who cofounded with me Datart, our startup. Thank you for this awesome year, with learned experiences that you can’t put a price on.

I thank the Brazilian agencies CAPES, CNPq and FAPESP for overall financial support of my research group and laboratory; and NVIDIA Corporation, for the donation of a Tesla K40 GPU used for this research.
“Do one thing a day.”

(João Menegola Sobrinho)
Abstract

From all skin cancers, melanoma represents just 1% of cases, but 75% of deaths\(^1\). Melanoma’s prognosis is good when detected early, but deteriorates fast as the disease progresses. Automated tools may play an essential role in providing timely screening, helping doctors focus on patients or lesions at risk. However, due to the disease’s characteristics — rarity, lethality, fast progression, and diagnosis subtlety — automated screening for melanoma is particularly challenging.

The objective of this work is to understand better how can we use Deep Learning — more precisely, Convolutional Neural Networks — to correctly classify images of skin lesions. This work is divided into two lines of investigation to achieve the objective. First, the study is focused on the transferability of features from pretrained CNN networks. The primary objective of that thread is to study how the transferred features behave in different schemas, aiming at generating better features to the classifier layer. Second, this study will also improve the classification metrics, which is the overall objective of this line of research.

On the transferability of features, we performed experiments to analyze how different transfer schemas would impact the overall Area Under the ROC Curve (AUC) training a CNN from scratch; transferring from pretrained CNN on general and specific image databases; performing a double transfer, in a sequence from general to specific and finally melanoma databases. From those experiments, we learned that transfer learning is a good practice, as is fine tuning. The results also suggest that deeper models lead to better results. We expected that transfer learning from a related task (in the case, from a retinopathy image database) would lead to better outcomes, but results showed the opposite, suggesting that adaptation from particular tasks poses specific challenges.

On the improvement of metrics, we discussed the winner pipeline used in the International Skin Imaging Collaboration (ISIC) Challenge 2017, reaching state-of-the-art results on melanoma classification with 87.4% AUC. The solution is based on the stacking/meta-learning from Inception v4 and Resnet101 models, fine tuning them while performing data augmentation on the train and test sets. Also, we compare different segmentation techniques - elementwise multiplication of the skin lesion image and its mask, and input the segmentation mask as a fourth channel - with a network trained without segmentation. The network with no segmentation is the one who performs better (96.0% AUC) against segmentation mask as a fourth channel (94.5% AUC).

We made available a reproducible reference implementation with all developed

\(^1\) American Cancer Society: cancer.org
source code for the contributions of this thesis.$^{2,3}$

**Keywords:** Neural Networks; Deep Learning; Melanoma; Convolutional Neural Networks.

---

$^2$ https://github.com/learningtitanis/melanoma-transfer

$^3$ https://github.com/learningtitanis/isbi2017-part3
Resumo

De todos os cânceres de pele, melanoma representa apenas 1% dos casos, mas 75% das mortes\(^4\). O prognóstico do melanoma é bom quando detectado cedo, mas deteriora rápido ao longo que a doença progride. Ferramentas automatizadas podem prover triagem mais rápida, ajudando médicos a focar em pacientes ou lesões de risco. As características da doença — raridade, letalidade, rápida progressão, e diagnóstico sutil — fazem a triagem de melanoma automática particularmente desafiadora.

O objetivo deste trabalho é melhor compreender como *Deep Learning* pode ser utilizado — mais precisamente, Redes Neurais Convolucionais — para classificar corretamente imagens de lesões de pele. Para isso, este trabalho está dividido em duas linhas de pesquisa. Primeiro, o estudo está focado na transferibilidade de características das redes CNN pré-treinadas. O objetivo principal desse tópico é estudar como as características transferidas se comportam em diferentes esquemas, com o objetivo de gerar melhores características para a camada de decisão. Em um segundo tópico, esse estudo incidirá na melhoria das métricas de classificação, que é o objetivo geral.

Sobre a transferibilidade das características, foram realizados experimentos para analisar a forma como os diferentes esquemas de transferência afetariam a Área sob a Curva ROC (AUC): treinar uma CNN a partir do zero; transferir o conhecimento de uma CNN pré-treinada com imagens gerais ou específicas; realizar uma transferência dupla, que é uma sequência de treinamento onde em um primeiro momento a rede é treinada com imagens gerais, em um segundo momento com as imagens específicas, e, finalmente, em um terceiro momento com as imagens de melanoma. A partir desses experimentos, aprendemos que a transferência de aprendizagem é uma boa prática, assim como é o ajuste fino. Os resultados também sugerem que modelos mais profundos conduzem a melhores resultados. Hipotetizamos que a transferência de aprendizagem de uma tarefa relacionada sob ponto de vista médico (no caso, a partir de um *dataset* de imagens de retinopatia) levaria a melhores resultados, especialmente no esquema de transferência dupla, mas os resultados mostraram o opposite, sugerindo que a adaptação de tarefas muito específicas representa desafios específicos.

Sobre a melhoria das métricas, discute-se o *pipeline* vencedor utilizado no International Skin Imaging Collaboration (ISIC) Challenge 2017, alcançando o estado da arte na classificação de melanoma com 87.4% AUC. A solução é baseada em *stacking/meta learning* dos modelos Inception v4 e Resnet101, realizando *fine tuning* enquanto executa a aumentação de dados nos conjuntos de treino e teste. Também compararam diferentes técnicas de segmentação — multiplicação elemento a elemento a elemento da imagem da lesão de pele e sua máscara de segmentação, e utilizar a máscara de segmentação como quarto canal — com

\(^4\) American Cancer Society: cancer.org
uma rede treinada sem segmentação. A rede sem segmentação é a que obteve melhor desempenho (96.0% AUC) contra a máscara de segmentação como quarto canal (94.5% AUC).

Nós também disponibilizamos uma implementação de referência reprodutível com todo o código desenvolvido para as contribuições desta dissertação.\textsuperscript{5,6}

**Palavras-chaves**: Redes Neurais; Aprendizado Profundo; Melanoma; Redes Convolucionais.
List of Figures

Figure 1 – Schematic representation of normal skin. Reproduced from (National Cancer Institute, 2018) ........................................ 16
Figure 2 – (a) Artificial neuron: a mathematical model of a biological neuron (b) Multilayer Perceptron (MLP) with one hidden layer ...................... 20
Figure 3 – Neurons of a convolutional layer connected to their receptive field. Reproduced from (KARPATHY, 2017) ................................. 21
Figure 4 – AlexNet architecture. Reproduced from (COLLET, 2017) .......... 22
Figure 5 – Residual learning building block. Reproduced from (HE et al., 2016) .................................................. 23
Figure 6 – (a) Inception module, naive version (b) Inception module with dimension reductions. Reproduced from (SZEGEDY et al., 2015a) .......... 24
Figure 7 – Transfer Learning strategy. A new SVM layer can be used to learn how to classify the features extracted. ......................... 26
Figure 8 – Augmentation of a dataset to reduce generalization error ........ 27
Figure 9 – Stacking strategy for classification ........................................ 28
Figure 10 – Comparison between clinical (left) and dermoscopic (right) images. Adapted from (BAR et al., 2012) ................................. 29
Figure 11 – Classical Computer Vision process versus Convolutional Neural Networks ................................................................. 35
Figure 12 – Samples for datasets used in the experiments on the transferability of features ................................................................. 36
Figure 13 – Size distributions of datasets employed on the experiments of transferability of features ......................................................... 38
Figure 14 – Images size distribution for all datasets. ................................. 44
Figure 15 – A visual panorama of our experiments performed on the improvement of metrics. ................................................................. 48
Figure 16 – Two ways to input segmentation information on the pre-trained model 53
List of Tables

Table 1 – A summary of literature of Deep Learning techniques applied on Melanoma Screening. Studies marked with asterisk use Transfer Learning from a pretrained network. .......................... 31

Table 2 – A summary of literature databases and metrics on Melanoma Screening. 32

Table 3 – Main results of the experiments on the transferability of features .... 39

Table 4 – Impact of the Deep Neural Networks (DNN) architecture choice. A deeper model (VGG-16) leads to best results, regardless of the experimental design. .................................................. 40

Table 5 – Results stratified by diagnosis difficulty of test images (Low, Medium or High), for VGG-M, transferring from ImageNet, with fine tuning. ... 40

Table 6 – Images and labels distribution across all datasets employed on the experiments performed at ISIC 2017 Challenge. .......................... 44

Table 7 – Official results on Private Leaderboard for Part 3 - Lesion Classification International Skin Imaging Collaboration (ISIC) Challenge 2017. ... 49

Table 8 – Official results on Private Leaderboard for subtask melanoma vs all on Part 3 - Lesion Classification on International Skin Imaging Collaboration (ISIC) Challenge 2017. ........................................ 50

Table 9 – Datasets used in the segmentation experiments. .......................... 52

Table 10 – Results for 3-channel segmentation ........................................ 54

Table 11 – Results for 4-channel segmentation ........................................ 54

Table 12 – During the competition, several hypothesis were tested to improve classification performance. This table summarizes what worked and what did not. ........................................... 58
List of Acronyms and Abbreviations

ANN Artificial Neural Networks. 18, 20, 21
AUC Area Under the ROC Curve. 39, 42, 43, 45, 47–50, 54
BCC Basal Cell Carcinomas. 16
CAD Computer Aided Diagnosis. 28, 50
CNN Convolutional Neural Networks. 17, 18, 21–24, 26, 27, 29, 33, 35, 42, 51, 52, 57
CV Computer Vision. 18
DA Data Augmentation. 27
DNN Deep Neural Networks. 17, 25, 26, 36, 37, 39, 40
GAN Generative Adversarial Networks. 59
GPU Graphical Processing Unit. 22
ILSVRC ImageNet Large Scale Visual Recognition Competition. 22
ISBI International Symposium on Biomedical Imaging. 18, 30, 36, 42
ISIC International Skin Imaging Collaboration. 18, 19, 28, 30, 31, 42, 43, 49, 51, 53, 59
MLP Multilayer Perceptron. 20
ResNet Residual Networks. 23, 28, 46, 48, 49
SCC Squamous Cell Carcinomas. 16
SVM Support Vector machine. 29, 36, 49, 50
# Contents

1 Introduction .................................................................................. 16  
1.1 Objectives .............................................................................. 17  
1.2 Outline .................................................................................... 18  
2 Literature Review ......................................................................... 20  
2.1 Artificial Neural Networks ....................................................... 20  
2.2 Convolutional Neural Networks .............................................. 21  
2.3 Deep Networks for Small Datasets ........................................... 25  
2.4 Melanoma ............................................................................... 28  
2.5 Discussion .............................................................................. 33  
3 On The Transferability of Features ............................................. 35  
3.1 Data and Methods ................................................................... 37  
3.2 Results and Discussion ............................................................ 39  
4 On The Improvement of Metrics ................................................ 42  
4.1 ISIC Challenge 2017 ............................................................... 42  
4.1.1 Data and Methods ............................................................. 43  
4.1.2 During the Competition .................................................... 46  
4.1.3 Results and Discussion ..................................................... 48  
4.2 Comparison of Segmentation Techniques for Classification Performance ........................................... 51  
4.2.1 Data and Methods ............................................................. 51  
4.2.2 Results and Discussion ..................................................... 53  
5 Conclusion .................................................................................. 56  
5.1 Lessons Learned ..................................................................... 56  
5.2 Open Questions and Future Work ........................................... 58  

Bibliography .................................................................................. 60
1 Introduction

This study aims to better understand Deep Learning’s application to medical datasets, focusing on computer-aided diagnosis for melanoma.

Skin cancer is the uncontrolled growth of abnormal skin cells. It occurs when unrepaired DNA damage to skin cells (often caused by ultraviolet radiation from sunshine or tanning beds) triggers mutations, or genetic defects, that lead the skin cells to multiply rapidly and form malignant tumors (PARKIN et al., 2011). It is a significant health problem, whose incidence has sharply risen in the past few years. In the 1930s, 1 in 1500 USA residents developed the disease; in the 2010s that incidence jumped to 1 in 59 (RIGEL, 2010)

There are different types of skin cancer:

- Basal Cell Carcinomas (BCC): abnormal uncontrolled growths that are in the skin’s basal cells, the deepest layer of the epidermis;
- Squamous Cell Carcinomas (SCC): abnormal uncontrolled growths that are in the skin’s squamous cells, the upper layer of the epidermis;
- Melanoma: abnormal uncontrolled growths that originate in the pigment-producing melanocytes, in the basal layer of the epidermis.

Figure 1 – Schematic representation of normal skin. Reproduced from (National Cancer Institute, 2018)
For 2018 alone, epidemiologists (American Cancer Society, 2018) predict 9320
deaths from melanoma on United States (one death every 57 minutes). Such high mortality
rate is due to how melanoma grows and spreads: it is among the most malignant cancers,
metastasizing (spreading to distant regions of the body via blood and lymphatic vessels)
in a short period.

Due to such malignancy, Melanoma early detection is imperative: if detected early,
it is among the most curable cancers (JERANT et al., 2000).

Melanoma can be diagnosed by simple visual examination of surface skin lesions,
but diagnosing it accurately is difficult, even for human experts (FRIEDMAN et al., 2008).
Medical personnel must invest in extensive training to become proficient at melanoma
diagnosis.

Early detection is hindered as the disease grows in incidence much faster than we
can train new doctors. Computer-aided skin cancer diagnosis appears as an opportunity
to alleviate such challenge. With the help of automated tools, primary-care professionals
can better decide whether or not to refer patients to the specialists, thus utilizing the
time of those scarce professionals more rationally (ESTEVA et al., 2017; CODELLA et
al., 2017b; FORNACIALI et al., 2016).

1.1 Objectives

From a Machine Learning point of view, melanoma diagnosis imposes the usual
challenges posed by medical images datasets: few training samples, class imbalance, and
hard-to-separate classes.

Deep Neural Networks (DNN), especially Convolutional Neural Networks (CNN),
are broadly used in image classifications tasks, usually outperforming traditional classifi-
cations techniques in competitions (LITJENS et al., 2018). Its main advantage is how it
can extract knowledge from large datasets, reaching outstanding results on datasets with
about a hundred thousand to millions of images. However, medical datasets usually have
few sample images, and for skin lesion datasets, it is no different.

The primary goal of this study is to understand better how deep architectures can
be used in the task of Melanoma Screening, since this field of research demonstrates that
these methodologies are recently increasing in popularity, by exploring different trans-
fer schemes to reuse knowledge from different tasks to improve our target task, and by
investigating the use of state-of-the-art techniques to improve melanoma classification
metrics.

The main objectives pursued in this research are the following:
• Push the boundaries of state-of-the-art melanoma classification metrics (performance) by applying Deep Learning techniques;

• Explore the impact of transferable features from a Deep Network trained in general (natural) and specific (medical) datasets;

• Explore the impact of Deep Learning procedures, such as fine tuning and data augmentation, on a small dataset.

Some hypotheses can be formulated to reach these objectives: (i) Transfer Learning is usually done with a pretrained network on natural images (like ImageNet). Transfer Learning pretrained on specialized images (like medical images) could provide small nuances natural images couldn’t provide; (ii) a Double Transfer scheme (when two knowledge transfers are done before the final transfer on the target task) could provide better features for melanoma images classification; (iii) using up-to-date techniques on image classification competitions could improve classification metrics.

Finally, the contributions of this work are:

• A better understand on the impact of transferable features from a Deep Network trained in general and specific (medical) datasets, with a paper published on IEEE International Symposium on Biomedical Imaging (ISBI);

• State-of-the-art results on automated Melanoma Screening, reaching first place on this task on ISIC Challenge 2017 (CODELLA et al., 2017a).

• We made available a reproducible reference implementation with all developed source code for both contributions.\(^1\)\(^2\)

1.2 Outline

Chapter 2 — Literature Review begins with an introduction on general topics related to Artificial Neural Networks (ANN) on Section 2.1. Section 2.2 discusses the background of CNN since its conception inspired on the biological process of vision to the most up do date architectures used in this work. Section 2.3 we motivate the use of Deep Learning for Computer Vision (CV), even for small datasets, where we present techniques to overcome the low data barrier. Finally, in Section 2.4 it is discussed the history of Melanoma Screening research, which provides us a better understanding of why the field pivoted from classical methods of Computer Vision to the most recent techniques of Deep Learning.

\(^1\) https://github.com/learningtitans/melanoma-transfer
\(^2\) https://github.com/learningtitans/isbi2017-part3
Chapter 3 — On The Transferability of Features discusses how we can take advantage of different Transfer Learning schemes to produce better features to a classifier. This chapter presents the results published in our conference paper entitled “Knowledge Transfer for Melanoma Screening with Deep Learning”. Section 3.1 discusses the datasets used, altogether with an explanation of the classification pipeline. Section 3.2 presents the results obtained as well as discusses them.

Chapter 4 — On The Improvement of Metrics details our winning submission for the ISIC Challenge 2017. Section 4.1.1 elaborates on the datasets and models used. Section 4.1.2 presents all hypothesis we wanted to test over the competition and what worked or not. Section 4.1.3 tells how we assembled all the models into one unique submission for the Private Leaderboard test set, as well as compare our submission with the top 5 competitors. Section 4.2 discusses the experiments performed on the improvement of metrics using several segmentation techniques.

Chapter 5 — Conclusion synthesizes our contributions, allowing a discussion about the impacts of this work in the context of skin cancer classification with Deep Neural Networks. It also provides guidelines for future work, considering the results obtained in this work can be further explored to understand transferability of features better and improve classification metrics.
2 Literature Review

2.1 Artificial Neural Networks

ANNs are machine-learning computational models, based on a large network of simple units called artificial neurons. A scheme of an artificial neuron can be seen in Figure 2a. An artificial neuron has \( N \) inputs and one output, the former representing the input Dendrites from a biological neuron, which receives the synapses from other neurons, and the latter representing the Axon, which sends the electrical activation signal to other neurons.

\[
\phi(x) = \phi(w \cdot x + b) = \phi(\sum_{i \in N} x_i w_i + b) \quad (2.1)
\]

\[
\min_w ||y - \hat{y}||^2 \quad (2.2)
\]

The artificial neuron model considers three main variables: inputs, weights, and biases. The output of a neuron is usually a non-linear function, applied to the sum of (i) an input vector \( \mathbf{x} \) with \( N \) dimensions, representing one input sample, multiplied by the weights vector \( \mathbf{w} \); and (ii) a bias term \( b \), conceiving the mathematical model described by Equation 2.1 (ROSENBLATT, 1958).

If a neuron is placed alongside other neurons, a layer of neurons is created. If two layers are created, and all outputs of the first layer are attached to all inputs of the second layer, a neural network is formed (Figure 2b). Given an input \( \mathbf{x} \) and an objective output \( y \),
one can use an optimization algorithm that finds the best weights and biases to minimize the square difference of a true output \( y \) and the network prediction \( \hat{y} \), causing the network to learn the data distribution in its input space.

## 2.2 Convolutional Neural Networks

The observation of biological processes is the starting point for many machine-learning algorithms. If one understands how a living being learns, a similar model can be developed and implemented in a machine. ANNs are inspired by how the brain works: different stimuli causes different activations on neurons. In Machine Learning, one could feed an ANN with each pixel intensity to each input neuron to classify an image, but that is just not how the visual cortex works. Hubel and Wiesel (HUBEL; WIESEL, 1968) identified two basic visual cell types in the brain:

**Simple cells** Their outputs are maximized by straight edges having particular orientations within their receptive field;

**Complex cells** They have larger receptive fields, and their outputs are insensitive to the exact position of the edges in the field.

Receptive fields are the region of visual space within visual stimuli that affect the firing of a single neuron. CNNs are similar to ANNs in a way that they are also made up of neurons, and that learn weights and biases. What makes CNNs suitable for image classification is the biologically inspired neurons arrangement: instead of layers of aligned neurons, CNNs are made of blocks of neurons. Each neuron on this block has a receptive field, a snapshot of a small region of the whole image, as it can be seen in Figure 3.

![Neurons of a convolutional layer connected to their receptive field](image)

Figure 3 – Neurons of a convolutional layer connected to their receptive field. Reproduced from (KARPATHY, 2017)

Fukushima (FUKUSHIMA, 1980) was the first to create a neural network, whose structure was similar to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel. A neural network was created with a cascade of three-layer blocks:
first, the input layer (the photoreceptor array); second, a layer with “S-Cells”, which shows characteristics to the simple cells proposed by Hubel and Wiesel, and a third layer with “C-Cells”, referring to the complex cells.

This idea introduces one of the leading advantages of CNNs over fully connected networks: local connectivity. By cascading layers of neurons with receptive fields, CNNs exploit spatially local correlation, enforcing a local connectivity pattern between neurons of adjacent layers. Thus, this architecture ensures that the first layers will have specific features (as edge detectors) and, as the network goes more in-depth, the last layers will have general (global) features.

In 1998, LeCun et al. (LECUN et al., 1998) created a pioneer CNN applied to handwritten digit recognition. The architecture was a 7-layer convolutional network, with three convolutional layers. At the time, the application on such networks had computational limits, thus they were not fully exploited. As technology on Graphical Processing Unit (GPU) advanced, CNNs were able to excel. By, parallelizing the training of 1.2 million images of a CNN with two GPUs, Krizhevsky et al. (KRIZHEVSKY et al., 2012) was able to achieve impressive results on a very complex task – the ImageNet Large Scale Visual Recognition Competition (ILSVRC) (DENG et al., 2009) – winning the competition with an incredible margin over classical image processing techniques. The architecture can be visualized in Figure 4.

![AlexNet Architecture](image)

Figure 4 – AlexNet architecture. Reproduced from (COLLET, 2017)

Multiple processing layers can learn representations of data with multiple levels of abstraction, yielding good results in complicated tasks (LECUN et al., 2015). Therefore, the Computer Vision community started to put effort on CNN architectures, building deeper models, increasing the level of abstraction.

Inspired by AlexNet, Simonyan et al. (SIMONYAN; ZISSERMAN, 2014) studied how depth increase could improve generalization on ImageNet dataset. The result was a significant model named VGG-16/19, because of their significant improvement on prior
CNN’s configurations by increasing the depth of the model to 16-19 layers. It is important to note the trade-off between building a more profound model and computational time: they used 4 GPUs to train the networks during a month. To exploit this trade-off, they also explored some shallower networks, known as VGG-S/M/F (CHATFIELD et al., 2014) (Slow, Medium, and Fast training), and their impact on performance.

The researchers showed that the deeper the model, the better the performance. The AlexNet and VGG networks are, nowadays, the top of mind layout of primary convolutional networks. They are a series of convolutional, pooling and activation layers followed by fully-connected layers which are used to solve a task. The first intuitive and straightforward idea one could suggest to improve performance using VGG-style networks would be to stack more layers. On the contrary, that does not improve performance. With the increasing depth, accuracy will get saturated and degrade rapidly. Overfitting does not cause that, and adding more layers would lead to a decrease in performance (HE et al., 2016). Why this behavior exists? What can we do to overpass this barrier?

Residual Networks (ResNet) (HE et al., 2016) were created with this problem in mind. Imagine a shallow(er) network constructed with convolutional layers. If you take this neural network and create a deeper counterpart with more stacked layers, one would hope that, at least, the training error would not decrease. We would hope that the network would learn how to perform an identity mapping from the top of the shallow network to the output of the deeper network. Experiments show that it is difficult for a network to learn an identity mapping by stacking nonlinear layers, and this is what has worsened the performance: direct mappings are hard to learn. Residual Networks propose a fix: instead of learning a desired underlying mapping from $x$ to $H(x)$, the network should learn a mapping created by the difference (residual) between them: $F(x) = H(x) - x$. So, instead of learning $H(x)$ as an identity, a convolutional block called Residual Block could be created, with the task to learn $F(x) + x$, as shown in Figure 5.

![Figure 5 – Residual learning building block. Reproduced from (HE et al., 2016)](image)

The authors hypothesize that it is easier to optimize this residual mapping than to optimize an unreferenced mapping. At the worst case, if the identity mapping should be the mapping to be learned, it would be easier to push the residual $F(x)$ to zero than to
fit an identity mapping by adding nonlinear layers. In conclusion, the residual block helps the network to rapidly fit an identity mapping if necessary, allowing the Machine Learning engineer to create a (much) deeper network. Another benefit from residual blocks is that they do not suffer much from vanishing gradients, caused by exponentially decreasing backward error signals, since they could easily travel back via the shortcut connections.

It is fascinating how tools like CNNs evolve over the years. When the computational cost was not such a big problem, this research field gathered efforts to learn how to solve complicated tasks like ImageNet better, leading from a very simple CNN like LeNet to a much deeper, high-performing network like ResNet. However, computational cost is a barrier when it comes to training a Convolutional Network. Inception networks (SZEGEDY et al., 2015b) seek not only a lower error rate by going deeper, but also scale up networks without increasing computational cost by going wider.

An Inception network is built with Inception Modules, created with two main aspects. First, an Inception Module tries to solve an information problem. There are numerous ways to create a convolutional layer in respect to the convolutional kernel. For example, a $5 \times 5$ convolutional kernel provides an output which is different from a $3 \times 3$ convolutional kernel, which provides something different from a $3 \times 3$ max-pooling kernel, and so on. However, which kernel provides the most meaningful information? An Inception model solves this problem by doing all transformations, and let the own network decide which is better (quite similar to boosting algorithms), resulting in a called “naive” version of an Inception Module, depicted in Figure 6(a).

![Inception Module](image)

(a)

This naive architecture increased information density, but is certainly not very computational-friendly, since we are increasing layers by putting them alongside, instead of stacking them. The naive version let the model decides which transformations are most informative at each Inception Module. Why not also let the model decides which filters are most informative? The key is the dimensionality reduction, which is done by performing $1 \times 1$ convolutions, resulting in the architecture portrayed at Figure 6(b). For example,
using 40 $1 \times 1$ filters, if the output of a previous layer has size $48 \times 48 \times 192$, these input maps will be compressed to $48 \times 48 \times 40$.

This architecture, known as Inception v1, also has evolved to a widely used architecture (including this work) Inception v4. The main difference between the four versions is the way convolutional layers are arranged inside an Inception Module, creating different architectures, which are almost always deeper and wider than the previous version.

From the acquaintance of specific neurons on a biological visual cortex, Convolu-
tional Neural Networks were developed showing that deep models can learn very general features on upper layers, yielding lower generalization error. Since these features are general, one could use the learned weights from a deep model to generate features from a database that the network has never seen — and they would be still meaningful. This idea (called Transfer Learning) is suitable for medical databases, which usually have very few image samples to train a CNN from scratch, and that is the basis for Melanoma research with Deep Learning.

### 2.3 Deep Networks for Small Datasets

DNNs are greedy for data and computational time. Stacking several convolutional layers will quickly increase the number of parameters to be trained, reaching the order of millions, even for relatively shallow networks. This fact is why CNNs work so well on image classification, because there is a lot of information capacity embedded on such networks. There is a need for a considerable amount of images to fill the network with information while training. But what if we don’t have this many images? To use DNNs on small datasets, some techniques need to be performed in order to achieve expected results. There are four which are the most widely used: Transfer Learning, fine tuning, data augmentation and stacking/ensemble techniques.

Traditional Machine Learning algorithms acquire knowledge from labeled or unla-
beled data (training data) with the goal to make predictions on new data (testing data). Usually, the new data is just like the data used for training. Transfer Learning, in contrast, allows the domains, tasks, and distributions used in training and testing to be different. The objective now is to train a model on a *source task* and apply the knowledge on a *target task*.

The strategy for Transfer Learning is freezing the weights of a pretrained deep neural network up to a chosen layer, usually until the second to last, and use this frozen network as a feature extractor, recycling knowledge from the first task to the new task. This way, a classifier layer can be used on the top of the network, learning how to classify the features extracted, as is shown in Figure 7.
Transfer Learning research has been around since 1995, with numerous proposed Transfer Learning schemes assuming that the source and target domains are related to each other in some sense. If this assumption does not hold, a negative transfer may happen, which may cause the learner to perform worse than no transferring at all (PAN; YANG, 2010).

Alex Krizhevsky showed in 2012 that CNNs could create generic features in upper layers, facilitating image classification and surpassing his opponents in ImageNet competition. However, are those features generic enough to perform Transfer Learning? Donahue et. al (DONAHUE et al., 2013) proposed evaluating whether deep convolutional activation features acquired from AlexNet could be re-purposed to different tasks. They showed that CNNs could learn features that have sufficient representational power and generalization to outperform sophisticated multi-kernel learning techniques with traditional hand-engineered features.

Yosinski et al. (YOSINSKI et al., 2014) explored the transferability of features from each layer of a DNN empirically, discovering three important aspects on knowledge transfer. First, transferability is negatively affected by two issues: fragile layer co-adaptation, when the transferred weights do not adapt to a subsequent randomly initialized convolutional layer, and performance drops due to representation specificity on upper layers. Second, there is a transferability gap as the distance between the source and target task increases, but they showed that even when the tasks are very different, it is better to transfer weights instead of randomly initialize them.

One last technique the authors showed that improve generalization performance
is Fine Tuning. Instead of initializing layers randomly and train from scratch, a training procedure should be performed over the transferred weights, with lower learning rate, by adding a fully connected layer with the number of outputs of the new task so that back-propagation can be performed. This way, the pretrained feature extractor will specialize the last layers, adapting the features to the new task. This technique is a crucial feature for image classification on small datasets, which is the case for skin lesion classification, since we do not have enough data to train from scratch, and the images are too specialized to have good performance with a pretrained network “off-the-shelf” on natural images like ImageNet.

For small datasets, we need to extract all the information we can get. When a picture of an object is taken, there are numerous factors (light, position, distance, etc.) that can change the visual perception of the object. What if we take a picture of an object from a different distance? What if we change the position of the object? What if we change the ambient light? It most certainly will change how the object looks like.

Figure 8 – The dataset can be augmented by performing some random transformations, so that the deep model would never see twice the same picture, improving generalization error. An image data generator was used to randomly apply different image transformations (rotation, width and height shift, shear, zoom and horizontal flip)

Since CNNs are used to perceive different patterns, if we perform arbitrary transformations on the images during training, we will be “augmenting” our dataset, creating new images from the same object like if the picture was taken in other conditions. This technique is called Data Augmentation (DA), and it is used to decrease generalization error. An example of Data Augmentation is illustrated in Figure 8. Altogether with Data Augmentation on training time, the technique can be used at test time. If the image is augmented at test time, we are changing how the network perceives the image, and so we are changing the probability of classification. Nowadays, it is common to perform Data Augmentation on test time, by providing to the trained network several augmentations of
the same test sample, and max-pooling or taking the average of the predictions on these augmented images.

Another useful technique that recently has grown in popularity in Machine Learning competitions is Stacking. Stacking is the act of piling different learners, to elaborate a more informed decision, as is depicted in Figure 9. For Melanoma Screening with the aid of CNNs, the different $d_L$ learners can be different architectures (VGG, ResNet, Inception), trained on different datasets. These learners’ opinions can be used to train another learner, a combiner $f()$, preferably nonlinear, which has the task to weigh the different opinions, maximizing the information.

![Figure 9 – Stacking is performed when a combiner $f()$, which is not restricted to being a linear combination as in voting, is used to aggregate opinions on different $d_L$ learners. Reproduced from (SANTINI, 2013)](image)

Finally, one could aggregate the opinions of several combiners $f_n()$ with another nonlinear learner $g()$, resulting in a technique called meta-learning. This way, the meta-learner will be weighing the combiners, also in a nonlinear fashion, to increase results. Stacking, meta-learning, and performing data augmentation in test time was crucial for our team to reach state-of-the-art results on ISIC 2017 Challenge melanoma classification task.

2.4 Melanoma

Melanoma Screening using a Computer Aided Diagnosis (CAD) system is challenging. Even for human specialists, it is difficult to diagnose melanoma (FRIEDMAN et al., 2008; ESTEVA et al., 2017), due to the similarity between malignant and benign moles. There are numerous melanoma databases, some are publicly available, but few have more than 2000 examples, making it difficult to train a CNN from scratch.

According to how they are acquired, medical skin-lesion images can be clinical and dermoscopic. Clinical images are macro-photographies taken directly from the patient. They only reveal the most superficial aspect of the lesion and suffer from reflexes and
other illumination challenges. Dermoscopic images are acquired with a dermatoscope, under controlled magnification and illumination, revealing more subtle features of the lesion. Those image types are exemplified in Figure 10. Dermoscopic images better reveal dots, globules, and networks, which are used to distinguish melanoma from other lesions.

![Figure 10 – Comparison between clinical (left) and dermoscopic (right) images. Adapted from (BAR et al., 2012)](image)

There is a lot of research in Melanoma Screening, but until 2015, the literature is mostly based on classical image processing techniques: preprocessing (image segmentation, hair, and occlusions removal), feature extraction (hand-crafted local and global features), and classification (non-linear classification algorithms) (FORNACIALI et al., 2016; PATHAN et al., 2018). With the success of the application of Deep Learning techniques in Computer Vision (KRIZHEVSKY et al., 2012) and the studies on Transfer Learning and Fine Tuning (DONAHUE et al., 2013; YOSINSKI et al., 2014), Melanoma Screening research pivoted to the use of Deep Learning models.

Masood et al. (MASOOD et al., 2015) were the first to publish a deep model on melanoma classification. The authors used a Deep Belief network trained from scratch in conjunction with a self-advising Support Vector machine (SVM) as a classifier. The quality of the model was tested on a minimal dataset (200 images for training and 100 images for testing), reaching 89% accuracy.

The first to introduce Transfer Learning as a technique to use CNNs on melanoma classification was Codella et al. (CODELLA et al., 2015). They extracted deep-activation features from a pre-trained CaffeNet model on the ImageNet dataset. They also compared the quality of the deep features with of the hand-crafted features. They showed that by using an ensemble of unweighted SVM score, averaging over all features (deep and hand-crafted) from 2624 images produces 93.1% accuracy on the test set.

After Codella, as it can be seen on Table 1, most of the researches on Melanoma Screening with deep models use a pre-trained network model to improve results (LIAO et al., 2016; SHOIEB et al., 2016; KAWAHARA et al., 2016; POMPONIU et al., 2016;
MAJTNÉR et al., 2016; SUN et al., 2016; FORNACIALI et al., 2016; CICERO et al., 2016; LOPEZ et al., 2017; MENEGOLA et al., 2017a; YU et al., 2017a; CODELLA et al., 2017b; VASCONCELOS; VASCONCELOS, 2017; ESTEVA et al., 2017). When the network is trained from scratch, it usually has low performance (NASR-ESFAHANI et al., 2016; YOSHIDA et al., 2016), or the dataset is too small (SABBAGHI et al., 2016), which could cause misleading results. There are two main aspects to note on Table 1. First, the literature is starting to employ deeper models on the research, ranging from AlexNet in early works (LIAO et al., 2016; SHOIEB et al., 2016; KAWAHARA et al., 2016; POMPONIU et al., 2016; MAJTNÉR et al., 2016), VGG (SUN et al., 2016; FORNACIALI et al., 2016; CICERO et al., 2016; LOPEZ et al., 2017; MENEGOLA et al., 2017a), ResNet (CICERO et al., 2016; YU et al., 2017a; CODELLA et al., 2017b) and, finally, Inception (VASCONCELOS; VASCONCELOS, 2017; ESTEVA et al., 2017), since recent results show that deeper models outperform shallower models (SUN et al., 2016; MENEGOLA et al., 2017a). The second aspect to note is how the ISIC archive is playing an important role on Melanoma Screening research, since the broad use of this publicly available database allows the community to compare and reproduce results. Also, since 2016, the ISBI has organized an annual competition, the ISIC Challenge on Skin Lesion Analysis Towards Melanoma Detection, featuring tasks as skin lesion segmentation and classification.

To achieve state-of-the-art results, Transfer Learning is just the first step. Four other techniques help to enhance best results: Fine Tuning, Data Augmentation, Lesion Segmentation, and Ensembles. There are two papers that compare models with and without Fine Tuning. First, our paper published on ISBI 2017 (MENEGOLA et al., 2017a), which will be discussed in more details on Chapter 3. Second, Sun et al. (SUN et al., 2016) did an extensive analysis of the use of CaffeNet and VGG, with and without Fine Tuning, using a database with 6584 clinical images. They extracted features from the second-to-last layer and used an SVM as a classifier. They also concluded that a deeper model (in this case, VGG) with Fine Tuning reaches the best accuracy (50.27% on 198 classes). On the same paper, they compared features extracted from deep visual features with hand-crafted features, also classified with SVM. They concluded that hand-crafted features give the best classification result on the task with 198 classes, with 52.19% accuracy. This result disagrees with Pomponiú et al. (POMPONIU et al., 2016), which used a pre-trained model to extract features from the last three layers, altogether with k-NN as a classifier, concluding that binary classification with deep visual features outperforms classification with hand-crafted features.

There are works (CODELLA et al., 2015; MAJTNÉR et al., 2016; CODELLA et al., 2017b) that show that hand-crafted features can be used to boost performance when used to perform an ensemble with deep models. Those works showed that SVM ensembles with deep visual and hand-crafted features outperform the case when those
Table 1 – A summary of literature of Deep Learning techniques applied on Melanoma Screening. Studies marked with asterisk use Transfer Learning from a pretrained network.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODELLA et al. (2015)*</td>
<td>Fusion of transfer CaffeNet features + hand-crafted features</td>
</tr>
<tr>
<td>MASOOD et al. (2015)</td>
<td>Deep Belief + SA-SVM</td>
</tr>
<tr>
<td>NASR-ESFAHANI et al. (2016)</td>
<td>CNN 2-conv + DA</td>
</tr>
<tr>
<td>SABBAGHI et al. (2016)</td>
<td>Deep auto-encoder + Bag-of-Features</td>
</tr>
<tr>
<td>YOSHIDA et al. (2016)</td>
<td>CNN 3-conv + DA</td>
</tr>
<tr>
<td>LIAO et al. (2016)*</td>
<td>AlexNet + FT</td>
</tr>
<tr>
<td>SHOIEB et al. (2016)*</td>
<td>AlexNet + Lesion segmentation</td>
</tr>
<tr>
<td>KAWAHARA et al. (2016)*</td>
<td>AlexNet + DA + Logistic Regression Classifier</td>
</tr>
<tr>
<td>POMPONIU et al. (2016)*</td>
<td>AlexNet + DA + kNN</td>
</tr>
<tr>
<td>MAJTER et al. (2016)*</td>
<td>Fusion of AlexNet + hand-crafted features + SVM</td>
</tr>
<tr>
<td>SUN et al. (2016)*</td>
<td>VGG16 + FT</td>
</tr>
<tr>
<td>FORNACLALI et al. (2016)*</td>
<td>VGG-M + FT + DA + SVM</td>
</tr>
<tr>
<td>CICERO et al. (2016)*</td>
<td>ResNet + FT + DA</td>
</tr>
<tr>
<td>LOPEZ et al. (2017)*</td>
<td>VGG + FT</td>
</tr>
<tr>
<td>MENEGOLA et al. (2017)*</td>
<td>VGG16 + FT + DA + SVM</td>
</tr>
<tr>
<td>YU et al. (2017)*</td>
<td>ResNet + FT + DA + Lesion Segmentation</td>
</tr>
<tr>
<td>CODELLA et al. (2017)*</td>
<td>SVM ensembles of (ResNet/U-Net/CaffeNet) + hand-coded features</td>
</tr>
<tr>
<td>VASCONCELOS; VASCONCELOS (2017)*</td>
<td>Inception v1 committee + FT + DA</td>
</tr>
<tr>
<td>ESTEVA et al. (2017)*</td>
<td>Inception v3 + FT + DA</td>
</tr>
<tr>
<td>RADHAKRISHNAN et al. (2017)</td>
<td>PatchNet: 7-layer CNN predicting on image patches as ensemble</td>
</tr>
<tr>
<td>GE et al. (2017)*</td>
<td>SVM classification from features extracted from ResNet50 + VGG16</td>
</tr>
<tr>
<td>YU et al. (2017)*</td>
<td>AlexNet + DA features encoded with Fisher Vector and classified with SVM</td>
</tr>
<tr>
<td>KWASIGROCH et al. (2017)*</td>
<td>Comparative study on VGG19, ResNet50 and VGG19 + SVM</td>
</tr>
<tr>
<td>GEORGAPOPOULOS et al. (2017)*</td>
<td>AlexNet + FT + DA</td>
</tr>
<tr>
<td>YU et al. (2017)*</td>
<td>ResNet50 + DA features encoded with Fisher Vector and classified with SVM</td>
</tr>
<tr>
<td>GE et al. (2017)</td>
<td>Joint VGG16 learning with Saliency Feature Learning</td>
</tr>
</tbody>
</table>

features are used separately. Also, Codella et al. (CODELLA et al., 2017b) used three different CNNs to extract features (ResNet50, CaffeNet, and U-Net) altogether with hand-crafted features. First, the skin lesion images were segmented with a U-Net network, which is broadly used on medical image segmentation. With a segmenting mask, the author created three different inputs: whole image (raw image), segmented image, and cropped image (bounding box around segmentation blob). Features were extracted from all types of input images, and finally, an ensemble with SVM is performed. The author reached 0.843 AUC on ISIC Challenge 2016 dataset, performing average ensemble over both whole images and crop over segmentation’s ground-truth images. This multi-context dataset (whole image and a cropped image) gives a significant boost in results.

The last technique literature shows that improve metrics is Data Augmenta-
Table 2 – A summary of literature databases and metrics on Melanoma Screening. Studies marked with asterisk use Transfer Learning from a pretrained network. FT: Fine Tuning | DA: Data Augmentation | acc: accuracy | map: mean average precision | auc: area under the ROC curve.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Database</th>
<th>Images</th>
<th>Dataset</th>
<th>Nº of Classes</th>
<th>Best Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODELLA et al. (2015)*</td>
<td>ISIC</td>
<td>Dermoscopic</td>
<td>334/2290</td>
<td>2</td>
<td>acc:93.1</td>
</tr>
<tr>
<td>MASOOD et al. (2015)</td>
<td>Sydney Melanoma Diagnostic Centre</td>
<td>Dermoscopic</td>
<td>120/170</td>
<td>2</td>
<td>acc:89.0</td>
</tr>
<tr>
<td>NASR-ESFAHANI et al. (2016)</td>
<td>UMCG</td>
<td>Clinical</td>
<td>70/100</td>
<td>2</td>
<td>acc:81.0</td>
</tr>
<tr>
<td>SABBAGHI et al. (2016)</td>
<td>National Institutes of Health</td>
<td>Dermoscopic</td>
<td>174/814</td>
<td>2</td>
<td>acc:95.0</td>
</tr>
<tr>
<td>YOSHIDA et al. (2016)</td>
<td>ISIC/Others</td>
<td>Dermoscopic</td>
<td>329/1431</td>
<td>2</td>
<td>auc:84.7</td>
</tr>
<tr>
<td>LIAO et al. (2016)*</td>
<td>DermQuest</td>
<td>Clinical</td>
<td>14799</td>
<td>23</td>
<td>map:70.0</td>
</tr>
<tr>
<td>SHOIEB et al. (2016)*</td>
<td>DermIS/DermQuest/DermNet</td>
<td>Clinical</td>
<td>337</td>
<td>2</td>
<td>acc:94.0</td>
</tr>
<tr>
<td>KAWAHARA et al. (2016)*</td>
<td>Dermofit</td>
<td>Clinical</td>
<td>1300</td>
<td>2</td>
<td>acc:94.8</td>
</tr>
<tr>
<td>POMPONIU et al. (2016)*</td>
<td>DermIS</td>
<td>Clinical/Dermoscopic</td>
<td>182/217</td>
<td>2</td>
<td>acc:94.0</td>
</tr>
<tr>
<td>MAJTNTER et al. (2016)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>auc:78.0</td>
</tr>
<tr>
<td>SUN et al. (2016)*</td>
<td>DermQuest</td>
<td>Clinical</td>
<td>6584</td>
<td>198</td>
<td>acc:50.2</td>
</tr>
<tr>
<td>FORNACIALI et al. (2016)*</td>
<td>EDRAS Interactive Atlas</td>
<td>Dermoscopic</td>
<td>364/907</td>
<td>2</td>
<td>auc:79.0</td>
</tr>
<tr>
<td>CICERO et al. (2016)*</td>
<td>Dermweb</td>
<td>Clinical/Dermoscopic</td>
<td>27963</td>
<td>24</td>
<td>acc:61.7</td>
</tr>
<tr>
<td>LOPEZ et al. (2017)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>acc:81.3</td>
</tr>
<tr>
<td>MENEGOLA et al. (2017)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>auc:83.5</td>
</tr>
<tr>
<td>YU et al. (2017)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>auc:80.4</td>
</tr>
<tr>
<td>CODELLA et al. (2016)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>auc:83.3</td>
</tr>
<tr>
<td>VASCONCELOS: VASCONCELOS (2017)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>map:64.3</td>
</tr>
<tr>
<td>ESTEVA et al. (2017)*</td>
<td>ISIC/Stanford</td>
<td>Clinical</td>
<td>129450</td>
<td>2</td>
<td>auc:94.0</td>
</tr>
<tr>
<td>RADHAKRISHNAN et al. (2017)</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>acc:76.8</td>
</tr>
<tr>
<td>GE et al. (2017)*</td>
<td>MoleMap NZ</td>
<td>Clinical/Dermoscopy</td>
<td>32195</td>
<td>15</td>
<td>acc:71.0</td>
</tr>
<tr>
<td>YU et al. (2017)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>auc:79.5</td>
</tr>
<tr>
<td>KWASIGROCH et al. (2017)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>auc:88.0</td>
</tr>
<tr>
<td>GEORGAKOPOULOS et al. (2017)*</td>
<td>Hospital of Wien</td>
<td>Dermoscopic</td>
<td>69/972</td>
<td>2</td>
<td>acc:92.3</td>
</tr>
<tr>
<td>YU et al. (2017)*</td>
<td>ISIC Challenge 2016</td>
<td>Dermoscopic</td>
<td>1279</td>
<td>2</td>
<td>auc:85.2</td>
</tr>
<tr>
<td>GE et al. (2017)</td>
<td>MoleMap NZ</td>
<td>Clinical/Dermoscopy</td>
<td>13292</td>
<td>15</td>
<td>acc:70.0</td>
</tr>
</tbody>
</table>
tion (VASCONCELOS; VASCONCELOS, 2017; YOSHIDA et al., 2016; ELMAHDY et al., 2017). Vasconcelos et. al (VASCONCELOS; VASCONCELOS, 2017) evaluated the impact of several Data Augmentation techniques over the ISIC Challenge 2016 dataset, like color and geometric transformations. Additionally, they also proposed a data augmentation technique based on the human specialist’s analysis, which uses the symmetry of the lesion to diagnose. They approximated the lesion segmentation ground truth by an ellipse, and augmented the image by changing the lesion’s axis size, but not its direction. By fine tuning a pre-trained GoogleNet (Inception v1) model for each Data Augmentation technique, they were able to compare each transformation independently, but also evaluate the impact of balanced and unbalanced datasets, as well as the committee of the average of the network’s individual classification. They stated that the committee of all models trained over a balanced dataset achieves the best result (0.643 MAP). In contrast with this result, motivated by how a human specialist analyzes a skin lesion, Yoshida et. al (YOSHIDA et al., 2016) proposes a preprocessing technique, with the same assumption as Vasconcelos’ – that the lesion symmetry should be used in favor to improve classification – by rotating the lesion over its major axis in a horizontal position. By using a CNN with 3 Convolutional layers, they stated that this proposed pre-process technique achieves better results (0.847 AUC) than flips (0.789) and rotation (0.826) over a dataset with 1760 images. These studies indicate that Data Augmentation improves metrics, but a thorough study should be performed on preprocessing techniques.

Esteva et al. (ESTEVA et al., 2017) used the most extensive dataset to the time of this writing in Melanoma Screening research. The dataset contains 129450 clinical images obtained from both ISIC and a non-public Stanford database. They used a pre-trained Inception v3 model, removed the decision layer and added a new decision layer with 757 classes, which they obtained after applying a disease partitioning algorithm to reduce the original 2032 disease classes to fewer, more independent set of diseases. After fine tuning the model, they performed cumulative sum over the decision layer’s probability output to calculate the probability for melanoma classification. They compared the specificity and sensitivity of over 20 human specialists with their model’s, concluding that the deep architecture employed outperforms the specialists, on average.

2.5 Discussion

Machine Learning has been greatly improved over the last few years. In Computer Vision, the need to write complex algorithms for image feature extraction was surpassed by algorithms who learns the feature extraction on their own. This kind of neural network evolved over the years in complexity, being able to realize long and heavy computations, allowed by the creation and price drop of GPUs. Nowadays, data plays one of the most
important roles: either you have and you get close to the solution of your task, or you
don’t, and you need to create ways to solve it. Improving Melanoma Screening metrics is
not the actual objective of this study, as Melanoma Screening is a use case for study on
small datasets. Esteva et al. (ESTEVA et al., 2017) showed it as a fact: with almost 130
thousand images, you can build a classification algorithm that is comparable to a human
specialist.

An important step to impact real world applications is the expansion and consoli-
dation of skin lesion databases. When this happens, maybe the field of Melanoma Screen-
ing research will not going to be about the study of Deep Learning on small datasets.
Instead, it might be the study of meaningful, relevant interpretation of skin lesion im-
gages, with the possibility to use a deep model to explain the image to a human specialist,
finally improving a human specialist’s diagnosis, overcoming unnecessary biopsies from
false positives and accelerating the treatment beginning for people in need.

Nowadays, it is still difficult to compare results between researchers (different
databases, different evaluation metrics, lack of reproducibility). The research on Melanoma
Screening needs to centralize its efforts. That is why the ISIC Challenge’s promotion and
expansion needs to be high priority in this field. ISIC Challenge 2016, the first edition of
the competition, brought together researchers to compete in three different tasks: Lesion
Segmentation, Feature Extraction and Lesion Classification. The winner at the time for
Lesion Classification, Yu et. al (YU et al., 2017a) used a ResNet with Fine Tuning and
Data Augmentation. At that year, there was no obligation for participants to write a
report of their solution, which turned out to be required on next year’s ISIC Challenge.
This obligation is constructive to this field of research, since this year we could compare
frameworks.

The competition’s lesion classification task had two subtasks: melanoma versus all,
and seborrheic keratosis versus all. From all 22 valid submissions on 2017 Challenge, only
one did not use CNNs (JIJI; RAJ, 2017), reaching 0.497 mean AUC over both classification
subtasks. Meaning that, not only classical image processing and classification techniques
are not suitable for this challenging task, but also that the community has pivoted to Deep
Learning techniques, as it has been seen on many other Computer Vision research fields.
The top 5 competitors (including our work, which will be discussed on Section 4.1) use a
pre-trained fine tuned CNN (MENEGOLA et al., 2017b; BI et al., 2017; MATSUNAGA et
al., 2017; GONZÁLEZ-DÍAZ, 2017; DEVRIES; RAMACHANDRAM, 2017), with some
pipeline differences, as some employ lesion segmentation, different types of networks and
ensembles.
3 On The Transferability of Features

From classical computer-vision point of view, it would be crucial to have a thorough understanding of the phenomenon that is being measured, so that the Machine Learning engineer would be able to elaborate a good preprocessing module and feature extractor, increasing the measurement metrics. For instance, in Melanoma Screening, the engineer would need to understand that a skin lesion specialist would be classifying the skin lesion based on the ABCDE rule, which are five signs that a skin lesion can be skin cancer. So he would learn that the “A” rule stands for asymmetry, and as a consequence, the engineer would need to elaborate a shape descriptor.

Nowadays, the whole classical Computer Vision pipeline, which applies numerous complex techniques, is being substituted by only self-learning image processing and feature extractor CNN, as depicted in Figure 11.

![Figure 11 – Classical Computer Vision process versus Convolutional Neural Networks.](image)

The objective in using CNNs is to substitute a whole lot of complex Computer Vision techniques for a unified, self-learning image processing and feature extractor.

From this perspective it may seem that CNNs would solve any problem. But there is no free lunch when it comes to CNN: to get a good feature extractor you need to feed the
network with (a lot of) data. When you don’t have enough, as is the case for Melanoma Screening, you need to transfer the weights from a pretrained network and use it as feature extractor. But how do you know the features being extracted are good enough to crack the task? Most importantly, what can you tune to provide better features? This chapter discusses this issue, by performing experiments that try to offer better insights in how can we extract the most from a transferred network, based on one of the main contributions of this dissertation: a conference paper accepted at the IEEE ISBI (MENEGOLA et al., 2017a) entitled “Knowledge Transfer for Melanoma Screening with Deep Learning”, from which some excerpts were taken to write this chapter.

DNNs are the state-of-the-art for image classification, but their use for medical images is challenging, since those models require extensive training sets (from dozens of thousands, up to several million images), and medical datasets are small. To bypass that difficulty, as stated in Section 2.4, most current literature employs Transfer Learning, ranging from simply using the output of the source DNNs as a feature vector, and training a completely new model (e.g., a SVM) for the target task; until using a pre-trained source DNN to initialize some of the weights of the target DNN, and training the latter as usual (YOSINSKI et al., 2014).

Because Transfer Learning seems so central for successful application of DNNs in Melanoma Screening, the main contribution of this work was studying it in more detail. Figure 12 illustrates both the source datasets (Retinopathy and ImageNet) and the target datasets used in our experiments. We attempt to answer the following questions: (i) What is the impact of Transfer Learning? Current art is already convinced it helps but by how much? (ii) Is it better transferring from a related (medical) but smaller dataset, from a more substantial unrelated (general) dataset, or from both? (iii) Does retraining the transferred network for the new task (known in the DNN literature as fine tuning) help and by how much?

Figure 12 – Samples from datasets used here (from left to right): Atlas (a); ISIC (b); Retinopathy (c); ImageNet (d). Each row shows a sample from a different class in the dataset. In this paper, datasets c and d are source datasets used for transferring knowledge to target models trained in the target task of Melanoma Screening, trained and evaluated in datasets a and b. Reproduced from (MENEGOLA et al., 2017a)
The dermoscopic examination is complicated and involves many types of lesions. There are other skin cancers, notably basal cell carcinomas, that can become a challenge for a classification model. Thus the question is: how should they be classified concerning other lesions? Another contribution is evaluating the impact of such decisions. We made available a reproducible reference implementation with all developed source code for this published paper\(^1\).

### 3.1 Data and Methods

The datasets employed to train and test the target models (Melanoma Screening) were:

**Interactive Atlas of Dermoscopy** (ARGENZIANO et al., 2002) a multimedia guide (Booklet + CD-ROM) designed for training medical personnel to diagnose skin lesions, containing 975 dermoscopic RGB images;

**ISIC Challenge 2016: Part 3** (GUTMAN et al., 2016) a subset of 1,279 dermoscopy RGB images (248 melanoma and 1031 nevus) from the International Skin Imaging Collaboration\(^2\).

The sources datasets employed for Transfer Learning (pre-training of the DNNs) were:

**Kaggle Challenge for Diabetic Retinopathy Detection**\(^3\) with a training set of 35126 high-resolution retina RGB images, with five severity classes: No Diabetic Retinopathy (25810 images), Mild (2443 images), Moderate (5292 images), Severe (873 images) and Proliferative DR (708 images);

**ImageNet Large Scale Visual Recognition Challenge 2012** (DENG et al., 2009) containing over 1 Million training RGB images labeled into 1,000 categories.

With a single exception, all protocols were based upon the VGG-M model proposed by Chatfield et al. (CHATFIELD et al., 2014). We also run a single comparison with the VGG-16 model (SIMONYAN; ZISSERMAN, 2014) to evaluate the impact of that deeper (and more expensive) architecture. Since we are trying to explore transfer learning schemes, and not trying to improve performance, these networks were chosen because their simplicity and their broad use on melanoma screening literature, as described on Chapter

---

\(^1\) [https://sites.google.com/site/robustmelanomascreening](https://sites.google.com/site/robustmelanomascreening)

\(^2\) [http://isdis.net/](http://isdis.net/)

\(^3\) [https://www.kaggle.com/c/diabetic-retinopathy-detection/data](https://www.kaggle.com/c/diabetic-retinopathy-detection/data)
Chapter 3. On The Transferability of Features

2. Because the original implementations are in MATLAB, which is not convenient for our computational infrastructure, we reimplemented those models in Lasagne\textsuperscript{4} and Nolearn\textsuperscript{5}.

In the experiments with Transfer Learning, we got the source networks pre-trained on ImageNet, trained them from scratch on Retinopathy, or fine-tuned on Retinopathy the model pre-trained on ImageNet. In the baseline (control) experiment without Transfer Learning, the networks were trained from scratch. In all networks, we ignored the output layer and employed an SVM classifier to make the decision. We did so for all experiments, including the fine-tuned ones, to avoid introducing extraneous variability. Whenever training was involved (when we fine-tune or train networks from scratch) we employed the technique of data augmentation. We randomly apply the following image transformations: zoom (\(\pm 30\%\)), rotation (\(0^\circ - 360^\circ\)), shear (factor \(\pm 30\)), height shift (\(\pm 30\) pixels), width shift (\(\pm 30\) pixels), and horizontal/vertical flip. To accomplish data balance, we only augmented the minority classes (Melanoma, Malignant, and Basal Cell Carcinoma, depending on the experimental design).

The size distributions of each dataset used in these experiments (except ImageNet) can be seen in Figure 13.

![Figure 13](image)

Figure 13 – The Atlas database is the most uniformly distributed dataset. The plots for Retinopathy and the ISIC Database took only the first 15 most common sizes, which comprehend more than 90\% of the databases. Both databases have high variation in size.

We evaluated three experimental designs, varying the labeling of the classes:

- Malignant \textit{vs.} Benign lesions: melanomas and basal cell carcinomas were considered positive cases (307 total images) and all other diagnoses were negative cases (668 images).

- Melanoma \textit{vs.} Benign lesions: melanomas were positive cases (256 images) while all other diagnoses were negative ones, removing basal cell carcinomas (668 images).

\textsuperscript{4} https://lasagne.readthedocs.io/en/latest/
\textsuperscript{5} http://nolearn.readthedocs.io/en/latest/lasagne.html
• Basal cell carcinoma vs. Melanoma vs. Benign lesions: here we have three classes, melanoma (256 images), basal cell carcinoma (42 images), and all other diagnoses under a single Benign label (668 images).

For all designs, we employed 5×2-fold cross-validation on the Atlas database. Our splits were semi-random, making an effort to balance as much as possible diagnose distributions, to avoid unnecessary variability (the datasets were stratified in respect to the labels). Our primary metric was the Area Under the ROC Curve (AUC); for the design with three classes, we computed three one-vs-one AUCs and reported their average.

3.2 Results and Discussion

As observed in Table 3, fine-tuning improves classification, either transferring from the small-but-related dataset (Retinopathy) or transferring from the large-but-unrelated task (ImageNet): that agrees with current literature in DNNs, which almost always endorses fine-tuning. Surprisingly, Transfer Learning from Retinopathy (also a medical-image classification task) led to worse results than transferring from the general task of ImageNet, even in combination with the latter. That might indicate that transferring from particular tasks poses unique challenges for overcoming the specialization — even if the source and target tasks are somewhat related. The best protocol we found was to transfer from ImageNet, with fine-tuning simply.

<table>
<thead>
<tr>
<th>Experimental Design</th>
<th>No Transfer</th>
<th>Retinopathy</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>no FT</td>
<td>with FT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>76.0</td>
<td>79.1</td>
</tr>
<tr>
<td>Malignant vs. Benign</td>
<td></td>
<td>72.8</td>
<td>82.5</td>
</tr>
<tr>
<td>Melanoma vs. Benign</td>
<td>75.7</td>
<td>76.0</td>
<td>77.9</td>
</tr>
<tr>
<td>Melanoma vs. Carcinoma vs. Benign</td>
<td>73.0</td>
<td>71.4</td>
<td>79.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72.8</td>
<td>80.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80.9</td>
<td>83.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>81.8</td>
</tr>
</tbody>
</table>

The experimental designs also showed differences in performance: in general, it was more natural to either group Basal cell carcinomas with Melanomas (Malignant vs. Benign), or to consider them as a separate class (Melanoma vs. Carcinomas vs. Benign), than to ignore them altogether (Melanoma vs. Benign). Those results suggest that organizing the labels affects the difficulty of the task, but the explanation might be just that Basal cell carcinomas are easier to diagnose than Melanomas.
Our results are consistent with current art on DNNs: Transfer Learning is a good idea, as is fine tuning. Our results also suggest, in line with literature in DNNs, that deeper models lead to better results. We expected that Transfer Learning from a related task (in our case, from Retinopathy, another medical classification task) would lead to better results, especially in the double transfer scheme, that had access to all information from ImageNet as well. The results showed the opposite, suggesting that adaptation from very specific — even if related — tasks poses specific challenges. Still, we believe that further investigation is needed.

The comparison between DNN architectures shows that — as usually observed for image classification — a deeper DNN performs better (Table 4). The results suggest that the experimental design is sensitive to the choice of lesions to compose the positive and negative classes, maybe due to the relative difficulty of identifying each of the types of cancer evaluated (Melanomas and Carcinomas).

Table 4 – Impact of the DNN architecture choice. A deeper model (VGG-16) leads to best results, regardless of the experimental design. All experiments with transfer from ImageNet and fine tuning.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Mal×Ben</th>
<th>Mela×Ben</th>
<th>Mela×Carc×Ben</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-M</td>
<td>82.5</td>
<td>80.9</td>
<td>83.6</td>
</tr>
<tr>
<td>VGG-16</td>
<td>83.8</td>
<td>83.5</td>
<td>84.5</td>
</tr>
</tbody>
</table>

We also show the results stratified by diagnosing difficulty (as indicated by the Atlas database metadata) in Table 5. Those results show that low-difficult lesions can mostly be solved by current state-of-the-art with relatively high confidence, while for difficult lesions performance is still little better than chance.

Table 5 – Results stratified by diagnosis difficulty of test images (Low, Medium or High), for VGG-M, transferring from ImageNet, with fine tuning. All: performance over the whole dataset. Low-, medium-, and high- difficulty cases represent respectively 38.1, 36.3, and 25.6% of the whole dataset.

<table>
<thead>
<tr>
<th>Experimental Design</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malignant vs. Benign</td>
<td>93.7</td>
<td>82.5</td>
<td>58.8</td>
<td>82.5</td>
</tr>
<tr>
<td>Melanoma vs. Benign</td>
<td>93.0</td>
<td>79.6</td>
<td>56.6</td>
<td>80.9</td>
</tr>
</tbody>
</table>

The results stratified by diagnosing difficulty suggest that current methods can already deal with the lower and middle spectrum of difficulty. On the other hand, challenging lesions appear really hard to diagnose. We believe that a referability framework (the easy cases are not referred to a doctor) could be potentially more fruitful than a diagnostics framework. Referring to the doctor both the cases in which the model has high confidence for the positive label and for the hard cases (for which the model has low
confidence), it might be more achievable in the short term than attempting to have high confidence for all cases.
4 On The Improvement of Metrics

Researchers from all around the world perform several experiments to understand better how their methods behave for the use cases they work on. That is was also what we did it on Chapter 3, where we were able to understand several Transfer Learning schemes and their impact on melanoma image classification. Researchers also perform experiments to improve classification metrics (performance), since the ultimate objective of the research on Melanoma Screening is to bring the research to the real world, consequently helping saving lives.

Those goals are reached while the community shares their ideas, acting as a catalyster to future works. As discussed on Section 2.4, the community has started to move from traditional techniques towards Deep Learning, following the general trend of Computer Vision, motivating our paper for ISBI 2017 (MENEGOLA et al., 2017a) and our team participation at Parts 1 and 3 of the ISIC Challenge 2017. We will discuss in this chapter our participation on Part 3, Lesion Classification, and our methods to reach the best AUC metric for melanoma classification.

As every other Machine Learning competition, the competitors aim to improve metrics no matter what. Looking at the competitors’ reports, we saw that, even though almost everyone used CNNs to perform Transfer Learning, there were many variations over the techniques used in each one’s solution. We made attempts on several techniques throughout the competition (as we will discuss on Section 4.1.1), and we figure out that most of the other competitors did. We noticed the lack of published works trying to measure the impact of the most common techniques used to improve classification metrics on Melanoma Screening, which motivated our research group to measure the statistical significance of most of these conventional techniques. One of those techniques is using segmentation to improve the metrics, and Section 4.2 discusses the experiments involving the choice of which segmentation technique would be used to look for its statistical significance.

4.1 ISIC Challenge 2017

Our team has worked on melanoma classification since early 2014 (FORNACIALI et al., 2014) and has employed Deep Learning with Transfer Learning for that task since 2015 (CARVALHO, 2015). Recently, the community has started to move from traditional techniques towards Deep Learning, following the general trend of Computer Vision (FORNACIALI et al., 2016). In 2017, the ISIC Challenge was posed to improve classification
metrics. This section describes our submission to this competition, described in our report entitled “RECOD Titans at ISIC Challenge 2017” (MENEGOLA et al., 2017b), from which some excerpts were taken to write this section.

4.1.1 Data and Methods

We aimed, from the start, at a Deep Learning solution. Our previous experience with the technique taught us that three significant bottlenecks would limit performance: the amount of training data, depth of the learning model, and availability of computational horsepower. Thus, we started by attempting to secure as much data and computational power as possible, to use models as deep as possible.

After those three significant issues were solved, there remains the excellent craftsmanship of optimizing the models. From the start, we aimed to get the highest possible rank at the challenge. If in (MENEGOLA et al., 2017a) we honestly stated that our aim was not pushing the envelope on model accuracy, here we can — also for the sake of honesty — state that we aimed to squeeze the last ounce of AUC from the models. Still, such AUC-squeezing goal was tempered by aesthetic considerations: we did not want an overly complicated solution. Added complexity had to bring proportional improvements in AUC, or we would prefer the more straightforward model.

The freedom to use external sources makes the number of training samples a critical factor: deep models crave for data. We collected several datasets to increase our training set. We restricted ourselves to publicly available (for free, or for a fee) reputable sources:

**ISIC 2017 Challenge** the official challenge dataset, with 2,000 dermoscopic images (374 melanomas, 254 seborrheic keratoses, and 1,372 benign nevi).

**ISIC Archive**

1 with over 13,000 dermoscopic images.

**Interactive Atlas of Dermoscopy** (ARGENZIANO et al., 2002) with 1,000+ clinical cases (270 melanomas, 49 seborrheic keratoses), each with at least two images: dermoscopic, and close-up clinical.

**Dermofit Image Library** (BALLERINI et al., 2013) with 1,300 images (76 melanomas, 257 seborrheic keratoses).

**IRMA Skin Lesion Dataset**

2 with 747 dermoscopic images (187 melanomas). This dataset is unlisted, but available under special request, and the signing of a license agreement.

---

1 The ISIC Archive: http://isdis.net/isic-project/

2 IRMA datasets: http://ganymed.imib.rwth-aachen.de/irma/datasets/
PH2 Dataset (MENDONÇA et al., 2013) with 200 dermoscopic images (40 melanomas).

Table 6 – Images and labels distribution across all datasets employed on the experiments performed at ISIC 2017 Challenge.

<table>
<thead>
<tr>
<th></th>
<th>ISIC 2017 Challenge</th>
<th>ISIC Archive</th>
<th>Atlas</th>
<th>Dermofit</th>
<th>IRMA</th>
<th>PH2</th>
<th>Total Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melanoma</td>
<td>374</td>
<td>701</td>
<td>500</td>
<td>76</td>
<td>40</td>
<td>187</td>
<td>1878</td>
</tr>
<tr>
<td>Keratosis</td>
<td>254</td>
<td>9</td>
<td>94</td>
<td>257</td>
<td>0</td>
<td>0</td>
<td>614</td>
</tr>
<tr>
<td>Nevus</td>
<td>1372</td>
<td>3335</td>
<td>1394</td>
<td>967</td>
<td>80</td>
<td>0</td>
<td>7148</td>
</tr>
<tr>
<td><strong>Total of Images</strong></td>
<td>2000</td>
<td>4045</td>
<td>1988</td>
<td>1300</td>
<td>120</td>
<td>187</td>
<td>9640</td>
</tr>
<tr>
<td><strong>deploy dataset</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>9640</td>
</tr>
<tr>
<td><strong>semi dataset</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7544</td>
</tr>
</tbody>
</table>

Figure 14 illustrates the images size distributions of all databases used in this chapter. The histograms for ISIC (Archive and Challenge) and Dermofit displays only the first 15 and 25 most probable sizes, respectively. It can be seen that these databases are the ones whose images most vary in size.

Figure 14 – Images size distribution for all datasets. The images from ISIC (Archive and Challenge) and Dermofit are the ones who most varies in size.

Our first strategy, following our experience with Deep Learning, was to compose a training set that was as large as possible. Thus, we took all available images from all datasets, except those that could cause annotation clashes with the challenge. We excluded the images without a diagnosis from the ISIC Archive, the ‘miscellaneous’ class from the Atlas, the images marked as ‘benign’ from IRMA, and the images marked as ‘atypical nevi’ from PH2. There was a suspiciously large cluster of benign-lesion images at
the ISIC Archive, all for 15-year old patients. Our validation numbers slightly improved after eliminating that cluster. We also found (near-)duplicates both inside and between the ISIC Challenge and the ISIC Archive datasets, thus, we created a procedure to avoid train–test contamination, ensuring that all (near-)duplicates stayed in the same (training or validation) split. After taking care of possible contaminations, the splits are obtained dividing the dataset with 85% of images for training and 15% of images for validation, stratifying by the labels (same proportion of labels on each set).

We called deploy the dataset assembling all six data sources, with the exclusions and deduplications explained above, resulting in 9,640 images. The images and labels distribution for this dataset can be seen on Table 6. Nevertheless, the best performance for melanoma on the official validation AUCs was obtained with a dataset that assembled just ISIC Challenge, ISIC Archive, and Interactive Atlas (with the restrictions explained above, and some additional exclusions). We called semi this subset of deploy with 7544 images. For keratosis, training with the full dataset (deploy) presented better results on the official validation AUCs.

Each training set will actually be separated into two splits: train and validation. The train split was used to find the weights in the Deep Learning models, and to train the SVM layers in the models that use it. The validation split was used to (a) compute what we called in the report internal validation AUC; (b) train the stacked SVM meta-model; (c) and — in a few cases — to establish an early-stopping procedure for the deep-learning training.

The results for melanoma raised challenging questions, since, for the internal validation AUCs, the larger dataset was better. The semi dataset could just be reflecting the official validation biases, or the semi is dataset better for melanoma. In the end, we decided to keep the results from models trained both in deploy and in semi for the aggregated decision (meta-learning).

Our previous experience (MENEGOLA et al., 2017a) showed that taking a model pre-trained on ImageNet, and fine-tuning it for skin lesion classification while performing data augmentation (Figure 8), is a sound strategy to get good results. It also showed that better models for ImageNet (usually deeper and more expensive) tend to be also better for the newly fine-tuned task. Therefore, we decided to concentrate our efforts in two models: ResNet-101 (HE et al., 2016) and Inception-v4 (SZEGEDY et al., 2016). Both are state of the art, available in multiple frameworks, and pre-trained for ImageNet with good results.

We initially trained independent models for melanoma and seborrheic keratosis, reflecting the structure of the challenge. However, we switched to a single 3-class model (melanoma, keratosis, and nevus) as we realized it would be prohibitive to carry twice the training and evaluation for the entire experimental plan.
4.1.2 During the Competition

After deciding which data and models we would use, we formulated hypothesis we would want to test:

1. Compare the baseline VGG-16 network to the deeper ResNet-101 or Inception-v4;
2. Compare standard-resolution images (224 for VGG and ResNet) to double-resolution images;
3. Contrast different strategies of class- and sample- weighting during training;
4. Compare normal training schedule with some form of curriculum learning (BENGIO et al., 2009; CHEN; GUPTA, 2015; SCHROFF et al., 2015);
5. Contrast different regimens of training and test augmentation;
6. Measure the impact of adding SVM as a final decision layer;
7. Use the patient data (age and sex) on classification;
8. Different model optimizers;
9. Add different types of per-sample normalization;
10. Add a final meta-decision.

Most of our attempts resulted in little to none improvement. We were not very diligent, however, in pursuing any factor whose effect size seemed small, we did not make all attempts in all models (ResNet and Inception) and all training sets (deploy or semi), and we performed no significance nor equivalence tests. The lack of statistical significance tests on the use of common techniques altogether with deep models of Melanoma Screening, motivated us to produce a paper (VALLE et al., 2017). Section 4.2 will discuss experiments on segmentation performed for the paper.

We sorted the factor list, placing first the biggest disappointments/surprises — the factors we most expected to improve the results but did not:

1. Image resolution – we tried both amending VGG-16 to accept larger inputs, and amending the augmentation procedure of Inception-v4 to accept larger images pre-cropping (but keeping the network input itself unchanged). Neither attempt improved the results;
2. Weighting – We attempted several class- and sample- weighting schemes, both to compensate the unbalancing of the classes, and the reliability of the annotations. The more complex the weighting scheme, the worse the AUCs — no weighting was the best weighting;

3. Validation and early stopping – we tried two ways to perform early stopping: first, when our internal validation AUC started to decrease, and second (more aggressive) when it refused to increase. With a single exception, there was no impact on the results;

4. Patient data – we attempted different encodings for incorporating the patient data (age and sex). The results were inconsistent, sometimes improving and sometimes worsening the results;

5. Curriculum learning – curriculum learning consists of scheduling training samples in order of difficulty (e.g., learning the easy cases first). The Interactive Atlas’ samples are annotated with a level of diagnosis difficulty (from a human point of view), allowing such scheme. We attempted a three-step schedule (starting with Atlas’ easy images, proceeding to Atlas’ easy and moderate images, and finalizing with all images). The results were worse than merely training with all images at once;

6. Segmentation information – Unfortunately, due to the time limitations, we were not able to incorporate the segmentation model learned in Part 1 in our classifier.

If most attempts disappointed, some were valuable, so we sorted the list by placing first the factors we believe helped the most:

1. Models + data – the mere transition to deeper models helped, but not by very much. It was the combination of deeper models and larger, more general datasets that boosted the numbers;

2. Data augmentation – for data augmentation, we randomly applied horizontal flip, color distortions (brightness, saturation, hue, and contrast) and crop. Train augmentation is not set to a fixed number of transformations: as long as the training persists, images are sampled from the training set, and random transformations are applied to them. We found out that test augmentation is critical as well: applying random transformations to the test sample, submitting those transformed samples to the network, and then pooling the results. When we employed an SVM decision layer after the network, augmentation was again fruitful; and when we stacked several models with a meta-learning SVM, augmentation was yet again important. We attempted several schemes for pooling, but a simple average pooling worked the best in all cases;
3. Per-image normalization – on Inception, normalizing the inputs to the network by subtracting the average image pixel improved results considerably. We did not have time to test this factor on ResNet;

4. Stacking models and meta-learning – a meta-learning scheme, using an additional SVM layer to learn the decision from the probabilities output by the models, gave the best results on the official validation AUC.

4.1.3 Results and Discussion

Figure 15 shows a subset of 48 out of more than a hundred models we evaluated. From the beginning, we noticed that the correlation between our internal validation AUCs and the official validation AUCs was far from perfect. In the plots shown, from left to right, the correlations are $R=0.58$, $R=0.77$, and $R=0.79$. The correlation was particularly bad for melanoma. That posed a challenge of choosing whom to trust: the official or the internal validation AUC. In the end, we chose to trust both (or neither) and included models that showed good performance in the two axes.

Another difficulty was that the best models for melanoma were not necessarily those for keratosis and vice-versa. We considered selecting different models for the different tasks, but in the end, we decided to pick the same set of models for both tasks and hope the meta-learning layer would make the adjustments. The meta-learning consisted in, for each sample, concatenating the decisions of each chosen model and using this as a feature.
Table 7 – Official results on Private Leaderboard for Part 3 - Lesion Classification on ISIC Challenge 2017. The teams are ranked on the mean AUC of two subtasks: melanoma vs all and seborrheic keratosis vs all.

<table>
<thead>
<tr>
<th>Rank</th>
<th>User</th>
<th>Organization</th>
<th>Mean AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kazuhisa Matsunaga</td>
<td>Casio and Shinshu University joint team</td>
<td>91.1</td>
</tr>
<tr>
<td>2</td>
<td>monty python</td>
<td>Multimedia Processing Group - Universidad Carlos III de Madrid</td>
<td>91.0</td>
</tr>
<tr>
<td>3</td>
<td>RECOD Titans</td>
<td>RECOD Titans / UNICAMP</td>
<td>90.8</td>
</tr>
<tr>
<td>4</td>
<td>Lei Bi</td>
<td>USYD-BMIT</td>
<td>89.6</td>
</tr>
<tr>
<td>5</td>
<td>Xulei Yang</td>
<td>Institute of High Performance Computing + National Skin Center, Singapore</td>
<td>88.6</td>
</tr>
</tbody>
</table>

vector for two binary SVMs (melanoma-vs-all, keratosis-vs-all). Those SVMs were trained using our internal validation set — thus we were prevented from evaluating them using the internal validation AUC. However, this scheme attained the best official validation AUCs. The submitted test run as well as our last official validation run were, thus, the result from a meta-model that assembled seven base models: three based on Inception trained on deploy; three based on Inception trained on semi; and one based on ResNets trained on semi. The results of those component models were stacked in a meta-learning layer based on an SVM trained on the validation set of deploy.

Table 7 displays the first five competitors on Private Leaderboard for Part 3 - Lesion Classification on ISIC Challenge 2017. An aspect to consider is how close the scores are between the first three contestants. The first place has a difference of 0.3% from our submission, with a considerably different model from ours.

MATSUNAGA et al. performed an ensemble of several ResNet50 models. They noticed that the task of classifying seborrheic keratosis was much easier than melanoma, so they integrated the certainty of the seborrheic classification model in the classification of melanoma, by creating a linear approximation of the models’ prediction outputs: if a sample is very likely to be seborrheic, it probably won’t be melanoma. But, as it can be seen in Table 8, which are the official results on Private Leaderboard for the subtask melanoma vs all only, they placed first on the competition because they had better results detecting seborrheic keratosis, and their linear approximation strategy may have increased their score on melanoma, but it did not help to outperform our model on this subtask. It is worth noticing that they used a considerably smaller dataset than our submission, which could have helped them to increase their metrics. Comparatively, regarding model complexity, their model seems to be less or equal than ours, since ResNet50 are much more shallow than our models, Inception v4 and ResNet101.
Table 8 – Official results on Private Leaderboard for subtask melanoma vs all on Part 3 - Lesion Classification on International Skin Imaging Collaboration (ISIC) Challenge 2017.

<table>
<thead>
<tr>
<th>Rank</th>
<th>User</th>
<th>Organization</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RECOD Titans</td>
<td>RECOD Titans / UNICAMP</td>
<td>87.4</td>
</tr>
<tr>
<td>2</td>
<td>Lei Bi</td>
<td>USYD-BMIT</td>
<td>87.0</td>
</tr>
<tr>
<td>3</td>
<td>Kazuhisa Matsunaga</td>
<td>Casio and Shinshu University joint team</td>
<td>86.8</td>
</tr>
<tr>
<td>4</td>
<td>monty python</td>
<td>Multimedia Processing Group - Universidad Carlos III de Madrid</td>
<td>85.6</td>
</tr>
<tr>
<td>5</td>
<td>T D</td>
<td>University of Guelph - MLRG</td>
<td>83.6</td>
</tr>
</tbody>
</table>

The competition organizers published an analysis on the competition’s results for all tasks (CODELLA et al., 2017a). The organizers presented the results on the top score by task and subtasks, while they discussed 4 major trends they observed:

1. The top scorers implemented ensemble techniques over Deep Learning models. All used additional data sources to train, either from ISIC (MATSUNAGA et al., 2017; MENEGOLA et al., 2017a), in-house annotations (GONZÁLEZ-DÍAZ, 2017), or external sources (MENEGOLA et al., 2017a).

2. Classification of seborrheic keratosis appears to be an easier task in this dataset, compared to melanoma classification, reflecting aspects of the disease, or bias in the dataset. The best performance on seborrheic keratosis came from the team that added additional weakly labeled pattern annotations to their training data (GONZÁLEZ-DÍAZ, 2017).

3. The top average performer was not the best in any single classification category.

4. The organizers performed simple methods of fusion (like average or a linear SVM) of all submitted probabilities on the test set. This method led to overall improvements in performance, reaching 92.6% AUC on the average of both subtasks (an increase of 1.5% over the top scorer), consistent with previous findings (MARCHETTI et al., 2017).

The main conclusion for these results is that Melanoma Screening research with CAD systems is still far from being used in real life applications. On the other hand, the competition motivates for future works, while being a benchmark for research on Melanoma Screening. We hope that this competition remains yearly since this would boost research in the field.
4.2 Comparison of Segmentation Techniques for Classification Performance

The ISIC Challenge 2017 not only united researchers from around the world, but also allowed researchers to inspect and learn from other researcher’s solutions. Analyzing competitor reports, we noticed the prevailing trend of CNNs usage, together with Transfer Learning, but essential distinctions on other aspects of the models, such as the use of segmentation techniques or how to perform an ensemble of the models.

We also did several attempts during the competition on specific hypothesis we believed it could improve classification metrics, but we were not very diligent in performing significance tests. With this in mind, we noticed the lack of published works trying to measure the impact of the most common techniques used to improve on Melanoma Screening, motivating our research group to measure the statistical significance of most of these standard techniques (like segmentation or data augmentation). Those results are about to published on the paper “Data, Depth, and Design: Learning Reliable Models for Melanoma Screening” (VALLE et al., 2017).

Since there are numerous ways to input segmentation information on the models, we explored a few segmentation techniques and applied the most promising one on the statistical significance test performed on the paper. This section describes the experiments performed on segmentation to improve classification metrics for automatic Melanoma Screening.

4.2.1 Data and Methods

We used mainly the same data sources we employed during the ISIC 2017 challenge, except by the IRMA Dataset, which was excluded because of the potential difficulties other researchers might face to obtain it. Even with that exclusion, the new dataset grew, due to a more careful matching of diagnostics among the sources (instead of dropping the doubtful cases).

To perform the experiments, the dataset was divided into train and test sets. The train set is composed of some images from ISIC Archive and ISIC Challenge 2017 datasets, while the test set is composed not only of other images from ISIC Archive but also of Dermofit and PH2 datasets. This test set simulates new data on the same dataset used for training, but also new data from different datasets, in a cross dataset fashion. All of these datasets provide the ground truth mask for the skin lesion segmentation, enabling us to train an automatic segmentation neural network.

In all experiments, we employed pre-trained models that proved successful for
Chapter 4. On The Improvement of Metrics

Table 9 – Datasets used in the segmentation experiments.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Melanoma</th>
<th>Keratosis</th>
<th>Nevus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISIC Archive</td>
<td>770</td>
<td>277</td>
<td>8392</td>
</tr>
<tr>
<td>ISIC Challenge 2017</td>
<td>374</td>
<td>254</td>
<td>1372</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISIC Archive</td>
<td>102</td>
<td>11</td>
<td>2998</td>
</tr>
<tr>
<td>Dermofit</td>
<td>76</td>
<td>257</td>
<td>331</td>
</tr>
<tr>
<td>PH2</td>
<td>38</td>
<td>0</td>
<td>160</td>
</tr>
</tbody>
</table>

the ImageNet task. To assess the impact of different segmentation techniques, we used the Inception-v4 (SZEGEDY *et al.*, 2016) architecture as a classifier, using the reference implementation available in TensorFlow/Slim v1.3.

Usually, there are two ways to perform segmentation to improve classification metrics. The first one is using a human segmented dataset, where a binary mask is created by a specialist to segment a skin lesion. This way is not very practical for real-life applications since we want to avoid the expensive use of an expert’s time, but it gives us a good baseline. The second way is to use a CNN to automatically segment a lesion, which is closer to real life applications. As we did in the ISIC Challenge 2017 (CODELLA *et al.*, 2017a), we used a segmentation network based on the work of Ronneberger *et al.* (RONNEBERGER *et al.*, 2015) to extract segmentation masks automatically. With the segmentation masks obtained from both methods (manually and automatically), we tried two approaches of inputting these segmentation masks:

- **Elementwise multiplication:** The binary mask, created manually by a human specialist or automatically by a CNN, is multiplied across all three channels of a sample, resulting in a 3-channel image.

- **Add the mask as a fourth channel:** instead of multiplying the mask across the image’s channels, we just add the segmentation mask as a fourth channel, hoping for the network to use all the information it can get.

For the experiments involving the segmentation mask as a fourth channel, we had to modify the Inception network, by adding adapter layers that receive four planes as input (the RGB planes and the segmentation mask) and output only three planes, as expected by the original networks. For this, we added three convolutional layers before the input, two layers with 32 filters, and a third with three filters. All convolutional layers used $3 \times 3$ kernels and stride of 1. Since Inception-v4 models require input images of $299 \times 299$ pixels, the adapter layer took $305 \times 305$-pixel images, to account for the two border pixels lost at each convolutional layer. Both ways of inputting the segmentation mask can be seen in Figure 16. All images (skin lesions and segmentation masks) were resized using Imagemagick’s resizing operation using its default Lanczos filter.
Figure 16 – Two ways to input segmentation information on the pre-trained model: (a) elementwise multiplication and (b) add segmentation as fourth channel with adapter layers.

In all experiments, we used the area under the Receiver Operating Characteristic curve (AUC) as main metric. Following the ISIC Challenge 2017, we use the mean AUC between the melanoma-vs-all and the keratosis-vs-all as the measured outcome in all experiments.

4.2.2 Results and Discussion

The objective of the paper (VALLE et al., 2017), was to analyze the statistical significance of each factor that could impact classification metrics, so the experiments were designed in a factorial fashion. We performed the experiments on segmentation in a similar way, but we did not do any statistical analysis on the results.

We trained four Inception-v4 networks in total, and for each of them, we used the pre-trained weights from ImageNet. One network was trained on the images without any segmentation, as a baseline. The three others were trained using some technique of segmentation: one network model was trained over the elementwise multiplication of the segmentation mask on the images. The two other networks were trained with the segmentation mask as a fourth channel, being one model trained with masks obtained
from a human specialist, and the other with an automatic segmentation using the U-Net model.

Since we cannot train a network with 3-channels images and test this trained network with 4-channels images, we performed full factorial experiments in each case. Table 10 summarizes the results for images with three channels. The trained models in each case (no segmentation and elementwise segmentation) were used to predict each other’s test set. From the Table, we can see that the best result was obtained without any technique of segmentation (0.960 AUC). We can also observe that in the case we train without segmentation and test with elementwise segmentation, the AUC dramatically drops (0.874 AUC), implying that the network trained without segmentation is using information from regions outside the segmentation mask to make the decision. This result is confirmed by analyzing the experiments where we train with elementwise segmentation. We can see that the metrics do not change too much (from 0.937 to 0.832 AUC), mostly because this model is now a specialist in looking only to the skin lesion.

Table 10 – Results for 3-channel segmentation

<table>
<thead>
<tr>
<th>Test</th>
<th>No segmentation</th>
<th>Elementwise Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No segmentation</td>
<td>0.960</td>
<td>0.874</td>
</tr>
<tr>
<td>Elementwise Segmentation</td>
<td>0.937</td>
<td>0.932</td>
</tr>
</tbody>
</table>

Table 11 summarizes the results for images with four channels. As we can see, the U-Net model is doing a good job segmenting the lesions, since the results in respect to a network trained with masks provided by a human specialist are quite similar. The model trained on a human specialist segmentation with four channels (0.945 AUC) is better than the model trained with elementwise segmentation (0.933 AUC). It indicates that the network trained with four channels uses information better than the model with three channels, despite the use of randomly initialized adapter layers to train a network with four channels.

Table 11 – Results for 4-channel segmentation

<table>
<thead>
<tr>
<th>Test</th>
<th>Human Specialist Segmentation</th>
<th>Automatic Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Specialist Segmentation</td>
<td>0.945</td>
<td>0.946</td>
</tr>
<tr>
<td>Automatic Segmentation</td>
<td>0.934</td>
<td>0.933</td>
</tr>
</tbody>
</table>

From those results, we can suppose that no segmentation is better than using segmentation alone. Even so, we decided to use the model trained with segmentation mask
as a fourth channel, as a factor to be confronted with the network without segmentation, so that we could measure the impact of segmentation with statistical significances. The results obtained in the work showed what we were observing on the results of these experiments, that introducing segmentation information to the model has a negative impact on classification metrics.
5 Conclusion

In this work, we approached the problem of using Deep Learning techniques in Melanoma Screening by exploring a better understanding of the transferability of pre-trained models’ features and the improvement of performance in Automated Melanoma Screening. In this chapter, we first outline the main contributions of this work, then provide directions for future works.

5.1 Lessons Learned

Literature on automated melanoma screening with Deep Learning is relatively new. The first papers were only published on 2015. For being such a new literature, there are gaps of knowledge that must be explored. This thesis tries to fill two gaps:

- How different techniques impact skin lesion classification, such as distinct transfer learning schemes, choice of lesions to compose different classes, diagnosing difficulty, and depth of network;
- Improve the classification performance.

Chapter 3 defines and discusses experiments that explores the first gap. The literature, as we can see on Chapter 2, either performs a network training from scratch or uses transfer learning with a network trained on ImageNet. We hypothesized that maybe the knowledge of a network trained on specific (diagnosing) and related (medical) task, could be more useful to classifying skin lesion. Also, we hypothesized that maybe transferring knowledge from an already transfered and fine tuned network (double transfer) could provide convenient information of a skin lesion image, by mixing knowledge from either a general or a specific task.

Exploring different transfer schemes led us to interesting results:

- The classification performance of a network trained from scratch and a network transfered from a specific task (retinopathy) with fine tuning is similar;
  - Performing fine tuning over random weights or weights trained over retinopathy leads to comparable results. Since we fixed the number of training epochs, the network is either being despecialized on the information gained from retinopathy or the retinopathy task does not provide useful information for skin lesion classification.
• The best classification performance is reached when transfer from ImageNet and fine tuning is performed;
  – The extensive and general ImageNet dataset used for training a deep network provides universal features, that really extracts useful information from images, helping to improve performance.
• Double transfer scheme actually has worse performance.
  – We believe this happens for the same reason training from scratch and fine tuning over retinopathy have similar results. Transferring to and from retinopathy may be introducing network despecialization.

Relative to the experiments regarding the choice of lesions to compose different classes, the best result is achieved when we are using three different classes, most likely because carcinoma may be an easier task for classification, providing a good performance, and since the AUC for three classes is being computed as the mean AUC of all three classes vs all, carcinoma is leveraging up the resulting AUC. This result shows that extra care is needed when choosing class groupings, and performance comparisons.

We also compared the training of the VGG-M network we used in the transfer learning schemes experiments and a deeper model, VGG-16, to explore the impact of the depth of the network. For this set up, we realized that a deeper model achieves best result.

Analyzing results stratified by diagnosing difficulty, we were able to see that a deep model performs well on easier diagnosis, and that, for real-world applications, instead of trying to diagnose skin lesions accurately, it could be better to build a referability system.

The literature on this field of research is relatively new, mostly due to the fact that now-famous CNN models are also recent. Additionally, looking closely to the deep models employed on the papers published on melanoma screening, we can observe that the literature is usually outdated in one or two years in respect to the state-of-the-art CNN models. For the ISIC Challenge 2017, we submitted a solution that uses Inception v4 models, aiming the use of up-to-date models. The combination of these really deep architectures with the union of several different skin lesion datasets, enabled us to achieve first place in melanoma classification task in the Challenge.

Besides the resulting contribution from state-of-the-art models and a large dataset, we also tried several other attempts of techniques, to improve classification performance. We listed and discussed all attempts on Chapter 4, which are also summarized on Table 12, pointing which ones worked, and which did not. Analyzing other competitor’s reports, we noticed a prevailing trend of CNNs usage, but they differ on other aspects of the models, such as the use of segmentation techniques or how to perform ensemble of models. That made us realize the need of a statistical significance study on which techniques impacts
Table 12 – During the competition, several hypothesis were tested to improve classification performance. This table summarizes what worked and what did not.

<table>
<thead>
<tr>
<th>ISIC Challenge 2017</th>
<th>What worked</th>
<th>What did not</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Models + data;</td>
<td>1. Image resolution;</td>
<td></td>
</tr>
<tr>
<td>2. Data augmentation (on train and test sets);</td>
<td>2. Weighting;</td>
<td></td>
</tr>
<tr>
<td>3. Per-image normalization;</td>
<td>3. Validation and early stopping;</td>
<td></td>
</tr>
<tr>
<td>4. Stacking models and meta-learning.</td>
<td>4. Patient data (sex, age);</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Curriculum learning;</td>
<td></td>
</tr>
</tbody>
</table>

positively on performance (VALLE et al., 2017). In that study, one of the experiments was to choose one of several segmentation techniques to be applied. The best model was achieved by adding adapter layers before the Inception input layer, enabling the use of segmentation mask as a fourth channel. Contrary to what we expected, we discovered that adding segmentation has a negative impact on classification metrics. Since the objective of this study was not perform an extensive research of all segmentation techniques in the literature, some techniques were set aside. For example, since we are performing pixel-wise multiplication of the mask and the lesion, we may be losing border information. Perhaps other techniques like cropping over the bounding box delimited by the lesion mask blob or even dilating it, could produce better results (CODELLA et al., 2017b).

All results regarding transferability, diagnosing difficulty, choice of classes and diagnosing difficulty were published in a conference paper (MENEGOLA et al., 2017a). Our research group believes that knowledge should go beyond reporting experimental design and results, and we are very proud to provide reproducible implementation with all the developed code for each of the contributions of this thesis.1,2

5.2 Open Questions and Future Work

Transferability of features Our experiments suggests that the double transfer scheme poses specific challenges to the network training since we were not able to measure why double transfer achieves worse results than simple transfer. It is also true for fine tuning on a simple transfer from a specialized database. Thus, one hypothesis

1 https://github.com/learningtitans/melanoma-transfer
2 https://github.com/learningtitans/isbi2017-part3
to be tested is if the network despecializes along the training, destroying possible
knowledge to be transferred. For future work, we will try to understand double
transfer better, focusing on the problem of network specialization/generalization.
To do so, the experiments published in the paper (MENEGOLA et al., 2017a)
should be reproduced, this time with really deep architectures (e.g., Inception v4),
with longer training schedules, larger databases, and a better way to define how to
perform training when using Transfer Learning.

**Improvement of the metrics** Besides the positive result in the ISIC Challenge 2017,
the classification metrics on automated Melanoma Screening are low for practical
purposes. Still, as observed on the challenge, the results on the test set were consid-
erably lower than on the training and validation set, suggesting that the model lacks
in generality over new data. We believe that more data would influence the improve-
ment of the metrics. As long as research on this field keeps growing, the datasets
on skin lesion images will grow as well, which will benefit this line of research. Our
study on Deep Learning with small datasets can be further explored. New kinds
of Data Augmentation could be explored, more specific to the model’s final task.
Experiments on the use of Generative Adversarial Networks (GAN) (GOODFEL-
LOW et al., 2014) could be performed, as this network could successfully generate
new samples for training data or use the knowledge obtained from the discriminator
network for Transfer Learning.


