UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Michel Silva Fornaciali

Reliable Automated Melanoma Screening for the Real World

Rastreio Confiável Automático de Melanoma para o Mundo Real

Campinas

2019

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Michel Silva Fornaciali

Reliable Automated Melanoma Screening for the Real World

Rastreio Confiável Automático de Melanoma para o Mundo Real

Thesis presented to the School of Electrical and Computer Engineering of the University of Campinas to obtain a Doctorate's degree in Electrical Engineering, in the area of Computer Engineering.

Tese apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas para obtenção do título de Doutor em Engenharia Elétrica, na área de Engenharia de Computação.

Supervisor: Prof. Dr. Eduardo Alves do Valle Junior

Co-supervisor: Profª. Drª. Sandra Eliza Fontes de Avila

Este exemplar corresponde à versão da tese apresentada à banca examinadora pelo aluno Michel Silva Fornaciali, sob orientação de Prof. Dr. Eduardo Alves do Valle Junior e Profª. Drª. Sandra Eliza Fontes de Avila

---

Campinas

2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

F767r

Fornaciali, Michel Silva, 1988-
Reliable automated melanoma screening for the real world / Michel Silva Fornaciali. – Campinas, SP : [s.n.], 2019.

Orientador: Eduardo Alves do Valle Junior.
Coorientador: Sandra Elisa Fontes de Avila.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Melanoma. 2. Aprendizado de máquina. 3. Imagem. 4. Redes neurais (Computação). I. Valle Júnior, Eduardo Alves do. II. Avila, Sandra Elisa Fontes de, 1982-. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Rastreio confiável automático de melanoma para o mundo real
Palavras-chave em inglês:
Melanoma
Machine learning
Image
Neural networks
Área de concentração: Engenharia de Computação
Titulação: Doutor em Engenharia Elétrica
Banca examinadora:
Eduardo Alves do Valle Junior [Orientador]
Ana Gabriela Salvio
Romis Ribeiro de Faissol Attux
Agma Juci Machado Traina
Anderson de Rezende Rocha
Data de defesa: 31-07-2019
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-0434-7020
- Currículo Lattes do autor: http://lattes.cnpq.br/5792124071429665

COMISSÃO JULGADORA – TESE DE DOUTORADO

Candidato: Michel Silva Fornaciali – **RA**: 071884

Data da Defesa: 31 de julho de 2019

Título da Tese: "*Reliable Automated Melanoma Screening for the Real World*"

Título da Tese em Português: "Rastreio Confiável Automático de Melanoma para o Mundo Real"

Prof. Dr. Eduardo Alves do Valle Júnior (orientador)

Dra. Ana Gabriela Salvio

Prof. Dr. Romis Ribeiro de Faissol Attux

Dra. Agma Juci Machado Traina

Prof. Dr. Anderson de Rezende Rocha

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Dr. Eduardo Valle. Eduardo is my mentor since my M.Sc., an example of a scientist and human being. He is responsible for my background as an independent researcher, continually guiding us to develop our abilities, believe in our capabilities and make the most of us. Also, Eduardo is my primary reference as a professor. I always had a crush for Education, and Eduardo taught me how to be a better teacher. I mirror myself and take inspiration from the way he prepares and conducts his classes.

As if it was not enough to have an outstanding supervisor, I was also fortunate to be co-oriented by another exemplary researcher: Prof. Dr. Sandra Avila. I also worked with Sandra since my M.Sc. She is another example to follow: she is always willing to help, passionate about what she does and strives to contribute to society through her research.

Thanks to Sandra and Eduardo, my postgraduate was an enjoyable period from which I have good memories. After these years, even though the mentor-student relationship has ended, it indeed remains the friendship with very dear and significant people in my career. It was an honor to work with both of you!

I also thank the RECOD and, in particular, the Titans research groups (I won't list names because they are so many!). You were present in the process, and this thesis is the result of the collaborative work we do.

Special thanks to my parents, Silvia and Daniel, for their encouragement and for believing in my dreams. They taught me that knowledge is my most valuable asset, and nobody can take it from me. Well, here I am thanks to your effort and support.

I thank my friends, and my boyfriend, Felipe Reis, for their patience and for the moments of relaxation, very important to maintain the balance of the journey. I also apologize to them for my impatience (sometimes) and (a little more frequent) absences.

Also, I thank UNICAMP, for the incredible years, unique opportunities, and many experiences since 2007 when I came to graduate in Computer Science.

I thank the development agencies (CNPq, CAPES, FAPESP) and companies (Microsoft Azure, NVIDIA) that collaborated directly or indirectly to carry out this research with funding or infrastructure investments. I gratefully acknowledge the donation of GPU hardware by NVIDIA Corporation.

Last but not least, I thank Google for recognizing and funding this research through a scholarship granted by the Latin American Research Awards (LARA) program, in the

years of 2016 & 2017.

Abstract

Melanoma outranks all skin cancers in fatalities, despite representing a minority of cases. Melanoma's prognosis, excellent when detected early, deteriorates fast if treatment is delayed, due to its tendency to metastasize. Early diagnosis is critical for a good outcome. Melanoma's screening must be a cheap, simple, and continuous process. Automated screening through image analysis may play an important role, especially in poor or isolated areas where the full-time presence of dermatologist is not feasible. Despite the recent advances in modern art of automated melanoma screening, much work is needed in order for automated solutions to be deployed in actual clinical settings. This work introduces an interdisciplinary view of automated melanoma detection. Besides advancing the machine-learning models, we aim at broader considerations that arise for using those models in the real world. We seek to address barriers — like lack of reproducibility, poor experimental designs, and lack of cooperation between disciplines — that prevent the development of reliable systems. From a computational viewpoint, we proposed an agenda for improving the reproducibility of existing art. Then, we delivered robust experimental designs to identify how deep learning-based approaches must be parametrized to recognize the disease better. We also investigated quality and bias issues regarding the datasets often used in this research area, raising questions about the reliability of automated tools. From a more legal viewpoint, we analyzed and criticized the current processes of regulating Artificial Intelligence-based medical devices. We showed the relationships between academic research and regulatory processes, highlighting the challenges and opportunities from the intersection of those topics. By cooperating with physicians, we studied the main characteristics of the disease, trying to incorporate them into the models, guiding the automated learning methods. We also briefly assessed physicians' perceptions of automated models over the years, identifying changes in relationships and typical concerns from specialists that should be taken into account when designing an automated system. Our main results relay in the computational view, promoting relevant methodological contributions for the community. We showed the importance of large, well-annotated datasets and powerful deep models to better extract the visual information from the images. We clarified the positive impact of transfer learning, and reinforced the need of data augmentation on testing (which although commonplace on other disciplines, was still not a norm on melanoma detection). Also, we proved that a lean deep-learning pipeline, with no pre-processing, no combination with classical feature extraction, and no post-processing or further decision layers is able to provide state-of-the-art performance. We showed that ensembles of those lean models are the best current alternative for obtaining top performance. Finally, we investigated the impact of dataset quality on the results, showing that the datasets typically employed have biases that can inflate the predictive power of the models, without

guaranteeing the generality of the method for other scenarios. Our findings and contributions enabled interesting foundations for future work, by promoting rigorous experimental protocols and analyses, by providing future researchers with valuable guidelines on how to design their models, and by bringing an interdisciplinary view for the problem. We hope those contributions will enable future advances to be more aligned with the needs of patients and medical workers, fostering the creation of automated screening methods for the real world.

Keywords: melanona; screening; real-world; deep learning.

Resumo

O melanoma supera todos os cânceres de pele em mortes, apesar de representar uma minoria de casos. O prognóstico do melanoma, excelente quando detectado precocemente, deteriora-se rapidamente se o tratamento for atrasado, devido à sua tendência a metástases. O diagnóstico precoce é fundamental para um bom resultado. A triagem do melanoma deve ser um processo barato, simples e contínuo. A triagem automatizada por meio da análise de imagens pode desempenhar um papel importante, especialmente em áreas pobres ou isoladas, onde a presença de dermatologista em tempo integral não é viável. Apesar dos recentes avanços na arte moderna da triagem automatizada de melanoma, é necessário muito trabalho para que soluções automatizadas sejam implantadas em ambientes clínicos reais. Este trabalho apresenta uma visão interdisciplinar da detecção automatizada de melanoma. Além de avançar nos modelos de aprendizado de máquina, buscamos considerações mais amplas que surgem para o uso desses modelos no mundo real. Procuramos abordar barreiras — como falta de reprodutibilidade, projetos experimentais ruins e falta de cooperação entre disciplinas — que impedem o desenvolvimento de sistemas confiáveis. Do ponto de vista computacional, propusemos uma agenda para melhorar a reprodutibilidade da arte existente. Em seguida, entregamos projetos experimentais robustos para identificar como as abordagens baseadas em aprendizado profundo devem ser parametrizadas para reconhecer melhor a doença. Também investigamos questões de qualidade e viés em relação aos conjuntos de dados frequentemente usados nesta área de pesquisa, levantando questões sobre a confiabilidade das ferramentas automatizadas. De um ponto de vista mais jurídico, analisamos e criticamos os processos atuais de regulação de dispositivos médicos baseados em Inteligência Artificial. Mostramos as relações entre pesquisa acadêmica e processos regulatórios, destacando os desafios e oportunidades a partir da interseção desses tópicos. Ao colaborar com os médicos, estudamos as principais características da doença, tentando incorporá-las aos modelos, orientando os métodos automatizados de aprendizagem. Também avaliamos brevemente as percepções dos médicos sobre modelos automatizados ao longo dos anos, identificando mudanças nos relacionamentos e preocupações típicas de especialistas que devem ser levados em consideração ao projetar um sistema automatizado. Nossos principais resultados residem no aspecto computacional, promovendo contribuições metodológicas relevantes para a comunidade. Mostramos a importância de conjuntos de dados grandes e bem anotados e de modelos profundos poderosos para extrair melhor as informações visuais das imagens. Esclarecemos o impacto positivo da transferência de aprendizado e reforçamos a necessidade de aumento de dados nos testes (que, embora comuns em outras disciplinas, ainda não eram uma norma na detecção de melanoma). Além disso, provamos que um pipeline enxuto de aprendizado profundo, sem pré-processamento, sem combinação com extração de

características clássica e sem pós-processamento ou camadas de decisão adicionais é capaz de fornecer desempenho de ponta. Mostramos que a combinação desses modelos enxutos é a melhor alternativa atual para obter o melhor desempenho. Finalmente, investigamos o impacto da qualidade do conjunto de dados nos resultados, mostrando que os conjuntos de dados normalmente empregados têm vieses que podem inflar o poder preditivo dos modelos, sem garantir a generalidade do método para outros cenários. Nossas descobertas e contribuições permitiram fundamentos interessantes para trabalhos futuros, promovendo protocolos e análises experimentais rigorosas, fornecendo aos futuros pesquisadores diretrizes valiosas sobre como projetar seus modelos e trazendo uma visão interdisciplinar para o problema. Esperamos que essas contribuições permitam que os avanços futuros sejam mais alinhados às necessidades de pacientes e profissionais da área médica, promovendo a criação de métodos de triagem automatizados para o mundo real.

Palavras-chaves: melanoma; rastreio; cenário real; aprendizagem profunda.

List of Figures

List of Tables

List of Acronyms and Abbreviations

AI Artificial Intelligence

ANN Artificial Neural Networks

AUC Area Under the ROC Curve

BoVW Bag-of-Visual-Words

CAD Computer-Aided Diagnosis

CBIR Content-based image retrieval

CNN Convolutional Neural Networks

DL Deep Learning

DNN Deep Neural Network

FDA United States Food and Drug Administration

GANs Generative Adversarial Networks

IEC International Electrotechnical Commission

IMDRF International Medical Device Regulators Forum

ISO International Organization for Standardization

SaMD Software as a Medical Device

SVM Support Vector Machine

Contents

1 Introduction

Melanoma outranks all skin cancers in fatalities, despite representing a minority of cases. The prognosis deteriorates as the disease progresses due to metastases [Tuong et al., 2012]. The number of new cases grows continuously: in the 1930s, 1 in 1 500 USA residents developed the disease; in the 2010s that incidence jumped to 1 in 59 [Rigel, 2010]. Melanoma is difficult to diagnose reliably (Figure 1), requiring extensively trained specialists. Unfortunately, the number of physicians trained to detect it does not grow proportionality [Voss et al., 2015]. Improving melanoma diagnosis is an urgent need.

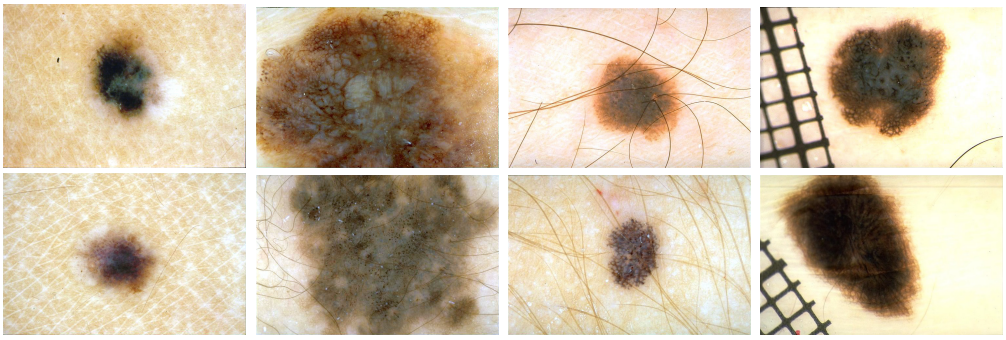


Figure 1 – Extracts of skin lesions from the Interactive Atlas of Dermoscopy dataset [Argenziano et al., 2002]. Melanomas (top row) looks similar to other skin lesions (bottom row), hindering diagnosis both for humans and machines. Image reproduced from Fornaciali [2015].

Medical training for melanoma detection relies on identifying typical patterns in skin lesions evidencing malignant cases. Searching for new cases must be a continuous process, especially in risk populations (Caucasian people in areas of intense sun exposure, or people with a personal or family history of the disease).

Although developed countries have the highest rates of melanoma, the incidence of new cases in developing countries, like Brazil, is a pressing issue to public or private health systems. Improving diagnosis and educating the population are critical actions to reduce fatalities.

Due to the visual appeal of the procedure to detect new cases, analysis of skin lesion images is an alternative to automated screening. Although automated screening does not replace doctors' examination, it brings several contributions: speeding up screening time, facilitating telemedicine programs, promoting triage of cases that requires specialist attention and enabling screening on poor or isolated areas where the presence of a full-time dermatologist is not feasible.

However, automated melanoma screening poses some challenges. The lack of annotated, high-quality data is by far the most severe. The difficulty in training machine

learning models to classify skin lesion images also deserves attention.

In the past thirty years, researchers proposed several approaches for automated melanoma screening (see Chapter 2). Companies also launched commercial products, being successfully used by dermatologists worldwide [Hand et al., 2015; Korotkov and Garcia, 2012]. Nevertheless, existing commercial solutions are expensive and do not provide completely automated diagnoses, but the freely available, fully automated academic solutions are not ready for real case usage.

Recent works claimed to achieve dermatologist-level performance for melanoma detection with Artificial Intelligence (AI) [Esteva et al., 2017; Haenssle et al., 2018; Tschandl et al., 2019]. In 2018, the United States Food and Drug Administration (U.S. FDA) approved the first fully AI-based device for healthcare: a system that detects diabetic retinopathy by eye images analyses [Food and Administration, 2018]. Brazilian efforts on the same subject also deserve attention [Pires et al., 2019].

Given the approval of regulatory agencies for full adoption of AI in medicine, and the existence of *apparently* technically sophisticated solutions, why fully automated solutions for melanoma detection have not been launched yet? It suggests the existence of gaps that we must overcome.

This research aims to reduce the barriers that prevent the adoption of automated solutions for melanoma screening. Our goal is to deliver technology and a process that could make benefits for the world.

1.1 Motivation

From the viewpoint of Computer Science, automated melanoma screening is a classical problem of image classification, a hot topic in the Computer Vision community. The related literature has evolved drastically since 2012, with the advance of Deep Learning techniques, increasing hit rates to levels never achieved [LeCun et al., 2015]. Deep learning has been bringing new perspectives to several research fields of medical imaging, like segmentation and disease identification [Litjens et al., 2017]. Those approaches also benefited automated melanoma screening.

Although the literature is improving the methods to classify skin lesion images, the existing solutions are still not robust enough for real-world scenarios.

Three main issues explain such fact: firstly, the **lack of standardization** that enables comparison and evaluation between different techniques; secondly, the **lack of annotated high-quality data** for the machine learning research; thirdly, the **scarcity of interactions between medical and computer science researchers** in this field. The last problem is the most severe because the interdisciplinary aspect of the research

requires the engagement and experience exchanging between the areas involved.

Here, we aim to address the three challenges, promoting an agenda to push forward the modern art of automated melanoma screening. We expect to initiate the development of a technology with the potential to transform the real world. Our team has worked on melanoma classification since early 2014 [Fornaciali et al., 2014], and has employed deep learning for that task since 2015 [Carvalho, 2015]. In this work, we expanded the scope to a multidisciplinary view of the problem. Although we will continue to contribute to the machine-learning view of melanoma screening, we will also address other aspects, such as studying ways of validating new technologies for health and investigating the issues of technology deployment for real usage.

The literature of automated melanoma screening evolved a lot during its thirty years old. In the late 1980s, it started with the fundamental question: *is it possible to analyze skin lesions automatically?* In the 1990s, some solutions emerged, not showing competitive results, but proving the concept.

In the 2000s, several solutions with different approaches populated the literature, advocating for advances in performance and adherence to medical protocols. The idea that seemed distant began to become closer to reality, raising the concern of physicians about the effectiveness of such methods. Literature, then, came to deal with another side of the problem: *are the existing products useful and safe?*

Nowadays, in the late 2010s, the promising results of automated skin lesion analysis leave no doubt about its value and potential. However, the problem is not resolved yet, and there is room for improvements. There are gaps in current solutions that prevent their full adoption by physicians and/or health professionals, like potentially inflated results, dubious validation protocols, and the existence of few solutions with recognition rates similar to the specialists [Esteva et al., 2017; Haenssle et al., 2018; Tschandl et al., 2019].

1.2 Objectives and Contributions

In this work, we aim to study *which are the design and evaluation aspects of automated skin lesion classifiers that must be improved to enable real-world usage.*

Our main contribution is an interdisciplinary approach, incorporating computational, medical, and legal aspects of the problem to deliver a skin lesion classifier to real-life usage. Moreover, other contributions are:

- we provide a broad survey of automated melanoma screening recent art based on deep learning techniques;

- we discuss the main pitfalls of current art, and we analyze the primary efforts for advancing these research field;
- we suggest legal aspects that should be followed by future works to promote automated melanoma screening for real-world scenarios;
- we contributed for the evolution of machine learning models for automated melanoma screening;
- we propose new methodological questions regarding automated melanoma screening that may guide future research.

1.3 Outline

We organized the remaining of the text as follows:

- **Chapter 2 — Literature Review**: this Chapter presents the existing solutions, analyzing past works and current art. We start describing techniques of image classification: although we focus on deep learning models, we briefly introduce other techniques such Bag-of-Visual-Words (BoVW) and Content-based image retrieval (CBIR). We also introduce the history of automated melanoma screening, resuming the main facts and findings of a thirty years timeline. Then, we summarize the current art based on deep learning models.
- **Chapter 3 — Critical Appraisal of Existing Art**: here we provide an interdisciplinary view of the current art, bringing the medical and legal aspects of designing an automated melanoma classifier. We also describe the main challenges — and on-going efforts — to promote automated melanoma screening for the real-world.
- **Chapter 4 — Advancing Machine Learning Models**: in this Chapter we describe our contributions towards reliable automated melanoma screening. We start introducing the materials (datasets) and methods (experimental approach) adopted in all activities developed during the Ph.D. Following, we list our main efforts, each one related with a major goal: improve machine learning models, propose robust experiment designs, create awareness of the main pitfalls inherent in the central problem. Such goals resume our methodological and experimental contributions. Each experiment synthesizes our hypotheses, objectives, the experimental proposals themselves, and the analysis of the results.
- **Chapter 5 — Conclusion**: we conclude the thesis putting our contributions into the perspective of the related art and future works. We discuss open questions, future directions, and new challenges that upcoming art will probably address. We end listing our publications and prizes.

2 Literature Review

Although automated skin lesion analysis seems to be a narrow topic, it is a vast research field. In this chapter, we survey the most relevant works and their contributions to the community.

We start our journey in Section 2.1 categorizing the study areas and their relationships. Then, we follow to Section 2.2 where we go through the most important existing past surveys, and existing gaps that motivated our survey. We also establish our focus: classification of skin lesion images. After that, we describe the technical aspects of image classification (Section 2.3).

We divide the central part of our survey into two sections: first, we present the history of automated melanoma screening, highlighting the main facts that contributed to shaping this research field as it is today (Section 2.4). Then, we dive into the modern approaches analyzing their construction from the computational point of view (Section 2.5).

We end our analyses in Section 2.6, recapitulating the main points and opening the discussions to the next Chapter, in which we provide an in-depth critique of existing literature. The technical interventions we made to answer some of the shortcomings we found in existing literature are presented in Chapter 4.

2.1 Study Areas

Korotkov and Garcia [2012] proposed a literature categorization, which we extend on Figure 2. In this work, we are not going to exhaust every single topic, but we are going to concentrate on the highlighted ones (and sub-topics, if any). Our main contribution is to analyze automated skin lesion analysis using the multiple views — medical, legal, and computing — necessary to promote automated tools in real-world scenarios.

The literature consists of disjoint publication areas/fields, with limited interaction: *Computer Science*, *Medicine*, and *Market* (or commercial products).

Medicine brings different types of publications, most of them being studies of relevant skin lesion cases (due to severity, rarity, or differentiated referral practices). That area is also responsible for publishing medical algorithms to facilitate skin lesion analyses, trying to identify malignant cases regarding determined criteria. The medical community also publishes works comparing/evaluating such algorithms in terms of efficiency. Those algorithms are relevant for our study because much of the existing automated art tries to reproduce, in software, the steps suggested by the medical rules. That brings to other important sub-topic of medical publications, that is the evaluation of Computer-Aided

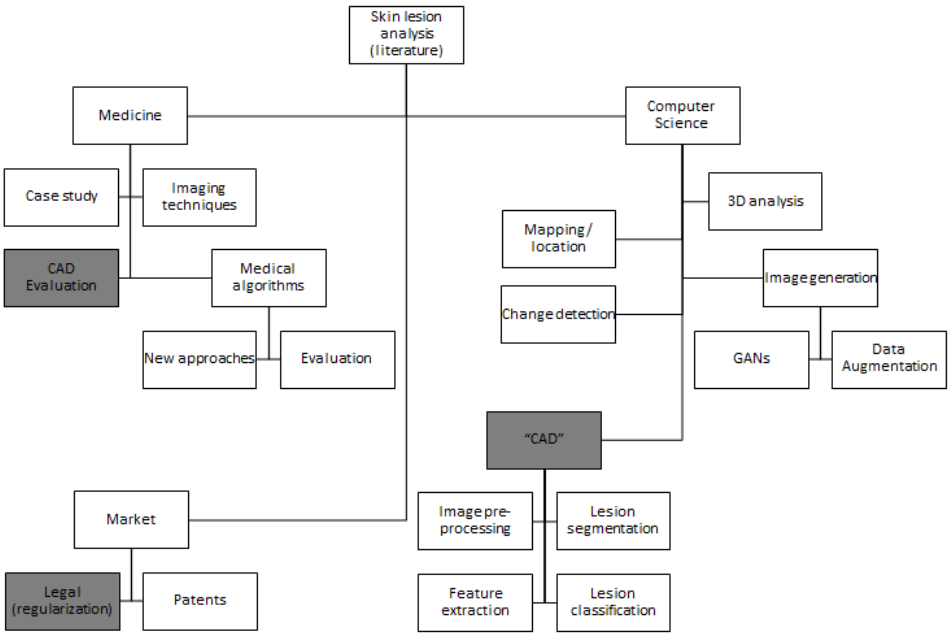


Figure 2 – Literature categorization tree. Image extended from Korotkov and Garcia [2012].

Diagnosis (CAD) systems: such works usually compare the effectiveness of the system with dermatologists or expose expert opinions on the use of such tools.

Computer Science brings several sub-topics related to skin lesion analysis, since image acquisition forms (like 3D, for example), until high-level image interpretation works (lesions change detection over time or lesion mapping across multiple images of the same patient). One of the most prominent sub-topics is the designing of CAD systems, which traditionally aims to reproduce, in software, the mentioned medical algorithms. As we will see later, that type of reproduction justifies the related topics linked to CAD systems: image pre-processing, lesion segmentation, feature extraction, and lesion classification. Recently, the advance of Deep Learning techniques promoted a race for medical data to enable the training of such complex computational models. As we know, new medical data is not easy to find/produce, so techniques for virtually generate such data became an essential topic in Computer Science research for skin lesion analysis [Bissoto et al., 2018b].

Finally, **Market** includes all publications related to commercial products. Here we have technical documents like patents and descriptions of the product legalization process together with the regulatory agencies. Depending on the potential risk offered by a medical device, it must be regularized before commercialization. CAD systems are typically one of them. The *Legal* literature relates to the standards, ISOs, and other official documents that guide medical devices regulation.

2.2 Analysis of the Literature

Automated skin lesion analysis is a vast, over thirty-year old, research field. Past surveys synthesize the advances of specific periods. However, such surveys often focus on only one — more rarely on two — of the contributing fields we identified above (Medicine, Computer Science, Market). The interdisciplinary aspect of skin lesion analysis is often overlooked.

Day and Barbour [2000] proposed the first survey of the related literature, summarizing the main findings. At that time, automated skin lesion analysis was far away to be used as a CAD system, but preliminary results indicated a promising bet for next years. Central issues delayed such promise. The authors argued that although the literature had more than ten years at the date, the research reporting limited the advances. Difficulties remained on three main factors: (a) lack of standardization on the testing sets, (b) lack of details describing proposed methods and (c) usage of small datasets to validate the models. As we are going to see, those issues are still open today, but new efforts are improving how the literature deal with them.

Twelve years later, Korotkov and Garcia [2012] revisited the literature describing the overall pipeline of skin lesion analysis in 4 steps: (a) image preprocessing, (b) lesion segmentation, (c) feature extraction, and (d) classification. They organized the information and trends of each step in several tables summarizing the current art, facilitating future research.

From 2012 till now, other surveys on automated skin lesion analysis were published. Masood and Ali Al-Jumaily [2013] provided a broad review for automated skin lesion analysis, including 31 works from 1993 to 2012. They proposed a general framework for assessing diagnostic models, with quality criteria for the following steps: calibration, preprocessing, segmentation, feature extraction, feature selection, training/testing separation, balancing of positive/negative classes, comparison of results, and cross-validation. In their conclusions, they highlighted the challenge of understanding and comparing existing art, promoting an agenda of standard benchmarks and validations, overlooked in previous studies.

Sathiya et al. [2014] presented several works on lesion classification, but since was published in 2014 does not cite any paper exploring modern techniques employed today. The summary of the literature was shallow and did not discuss problems or future directions for the current art. Fernandes et al. [2016] also suffer from the same problem, even being published two years later. Their comparison lies between poor techniques of color constancy and skin lesion analysis, not exploring the potential of works employing advanced BoVWs and DL models. Their conclusions were directed towards a single dataset, not analyzing whether the conclusions generalized for other data. Finally, Oliveira et al.

[2016] also reviewed the literature of skin lesion classification, but limited their analysis in local and global patterns (color and texture descriptors) and traditional classification techniques (Artificial Neural Networks (ANNs), Support Vector Machine (SVM), Bayesian Networks, Decision Trees, etc.). They did not even mention modern works based on Deep Learning.

Considering the shortcomings of the last surveys, we also did a review of the art in 2016 [Fornaciali et al., 2016]. We analyzed the literature under the same pipeline proposed by Korotkov and Garcia [2012], and contrast the problems listed by Day and Barbour [2000] in light of the current literature. In our work, we also explored the computational aspects of Deep Learning, describing the early works of the time and comparing their performance in practice with those of classical solutions. We concluded that Deep Learning would definitely be the new bet in terms of computational models for the analysis of skin lesions and that the typical problems of the literature, such as lack of data and difficulty in reproducing work, were still present, although the first initiatives to mitigate them.

As literature was heavily based on classical techniques, its review took time to be entirely focused on modern techniques. Pathan et al. [2018] provided a broad review of each step of the pipeline for automated skin lesion analysis, as also included modern approaches based upon deep learning techniques. Nevertheless, their coverage on that matter was shallow, not including recent papers and not analyzing technical aspects of deep learning approaches.

As far as we know, Brinker et al. [2018] proposed the first review of contemporary art focused on deep learning approaches. They provided a systematic review, only including works dealing with skin lesion classification that presents their solution understandably, discussing the results sufficiently. Their findings point out that current deep learning approaches for automated melanoma screening usually train their methods from scratch, or use deep learning only for feature extraction or employ transferring learning methods. We will discuss such terms and approaches in Sections 2.3 and 2.5. They also corroborate with our previous findings ([Fornaciali et al., 2016]) on the effectiveness of deep learning and the ongoing challenges to improve the accuracy rates of automatic methods.

A shortcoming of the surveys cited above is an exclusive focus on the computing aspects — with medical issues mentioned in passing, if at all. Lisboa and Taktak [2006] was the first paper to discuss ethical and legal issues on a computer system for medicine. They surveyed the literature showing how AI has gained more space in the medical area, relating the most employed algorithms, methodologies for model selection, and exploring the need for rigorous result evaluation. That survey was a pioneer in bringing a multidisciplinary view to AI in Medical Science, with a broad viewpoint. Here, we will focus on automated skin lesion analysis.

We focus on deep learning solutions for automated melanoma screening. We take the relay of recent surveys, with some overlap, focusing on papers from 2015 onwards, when the literature started to adopt deep learning techniques. For a broad perspective of the prior art, we refer the reader to our first survey [Fornaciali et al., 2016]. This work includes medical perspectives of the problem and, as Lisboa and Taktak [2006], we also discuss legal aspects of using a CAD system. We aim to understand how such three aspects are related and should be addressed together to enable automated skin lesion analysis for the real-world.

Such interdisciplinary initiative is pioneer, and we expect the benefits of this work to have long-ranging impacts, both scientifically and socially.

2.3 Image Classification

From the viewpoint of computer science, automated melanoma screening with images is a problem of image classification (see Figure 3). Here we briefly describe the most popular approaches for image classification. For further information, we refer the reader to our survey [Fornaciali et al., 2016], from which we reproduce part of the text to compose this Section.

To solve image classification problems, we use two sets of data, one to train the model and the other to test it. The *training set* goes through a feature extraction routine that describes the images mathematically. Those features and image labels are used to train a classifier that generates a predictive model. When the predictive model is presented to other images (the *testing set*) — of which we do not know the labels — we extract features in the same way as in training step. Then, the model infers to which classes the new images belong. Here we use the concept of *supervised learning*, in which we know from which class each image of the training set belongs. In *unsupervised learning*, we do not know the classes of the training images, and the data are grouped according to the similarity of their characteristics.

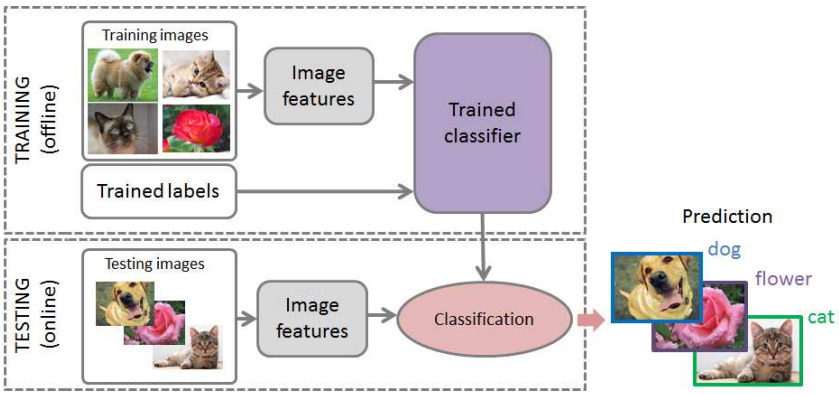


Figure 3 – Classical representation of an image classification system. Image reproduced from Fornaciali [2015].

In the past decade, literature of image classification focused on two models: BoVWs and DNNs. BoVWs and DNNs follow the same information-processing pattern: they extract from the pixels progressively higher-level features until they can establish the image label. However, they make entirely different choices to implement that pattern: BoVWs are shallow architectures, with three layers of feature representations (low-level local features, mid-level global features, high-level classifier); DNNs have many layers: from a dozen up to hundreds. BoVWs estimate/learn parameters conservatively: no learning at low level, none to a little unsupervised learning at mid-level, and supervised learning only at the high-level classifier. DNNs aggressively learn millions of parameters throughout the entire model.

Another crucial difference is that the BoVW approaches have a clear separation between the stages of image processing, while in DNN approaches such separation is abstracted, often suppressed. In BoVW-based CAD systems for melanoma, the images pass through 4 steps (Figure 4): (a) image preprocessing, (b) lesion segmentation, (c) feature extraction, and (d) classification. Each of these steps has its particularities, challenges, and solutions.

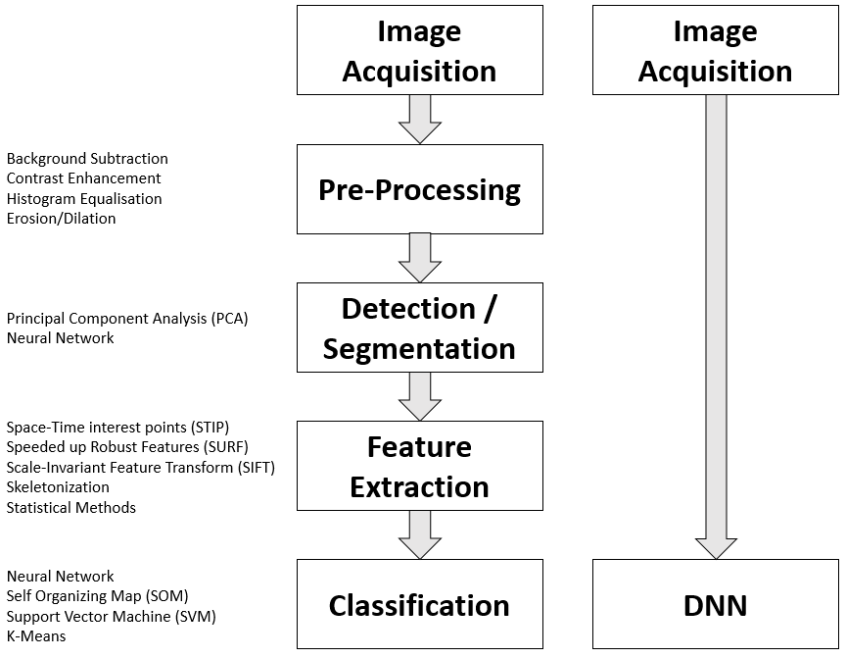


Figure 4 – Comparison of the Classical Computer Vision Approach versus The Deep Neural Networks model. The objective in using DNNs is to substitute a whole lot of complex Computer Vision techniques for a unified, self-learning image processing and feature extractor. Image adapted from Menegola et al. [2018].

Here we are interested in the classification step with DNNs, which usually suppress previous steps. For the readers interested in the other steps, we recommend the following papers: Celebi et al. [2015] summarize the main advances for *lesion border detection*; *feature extraction* generally lies on asymmetry and color information of the lesions, which are respectively discussed by Premaladha and Ravichandran [2014] and Madooei and

Drew [2016]. *Preprocessing* approaches are discussed in all the cited papers. For a complete tabulation of materials and methods employed in the four steps, we also recommend the works of Korotkov and Garcia [2012]; Pathan et al. [2018], but they focus on traditional approaches for image classification. For a broad review of such tasks and medical fields besides skin lesions, the survey of Litjens et al. [2017] is a good starting point.

Deep architectures have a long history compared to BoVW approaches, if we accept as deep the model presented by Fukushima [1980]. However, the term *Deep Learning* only caught on after the work of Hinton et al. [2006]. They published their work at a moment ripe for the aggressive Deep Neural Network (DNN) models to thrive, with the availability of large training datasets, and powerful many-core GPU processors. In DNNs, feature extraction and label classification are seamless: deep networks *learn* to extract the features.

The definitive victory of DNNs in Computer Vision came in 2012, when Krizhevsky et al. [2012] won the classification challenge of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012. They employed a large, deep Convolutional Neural Networks (CNN) , considerably improving the art at the time.

Although DNNs have now almost completely supplanted BoVWs, medical applications, where the training datasets are comparatively tiny, are very challenging for DNNs. For example, consider the ImageNet/ILSVRC competition, which launched DNNs: it has 1.2 million images in the training set; meanwhile, most reported melanoma datasets have less than ten thousand images. Nevertheless, with the technique of transfer learning, DNNs become competitive even in the scenarios of few data. We will discuss that on next Section.

2.3.1 Deep Learning

Deep Learning for visual tasks usually involves seamlessly learning every step of the classification process, from feature extraction to label assignment. That pervasive learning improves generalization, but brings its own challenges: DNNs estimate a huge number of parameters, requiring large amounts of annotated data and computational resources.

DNNs are based on Artificial Neural Networks, which have a long history, the first ideas dating back to 1943, when McCulloch and Pitts [1943] proposed a mathematical model for the biological neuron. However — as we did with the "words" in "visual words", we caution the reader against the "neural" in "neural networks". The similarity of modern DNNs to biologically plausible neural models is questionable, to say the least. Today, the justification for DNNs rests not on biological metaphors, but on their — empirically validated — amazing performance in a broad variety of tasks.

DNNs may also be approached as multi-layered generalized linear models learned

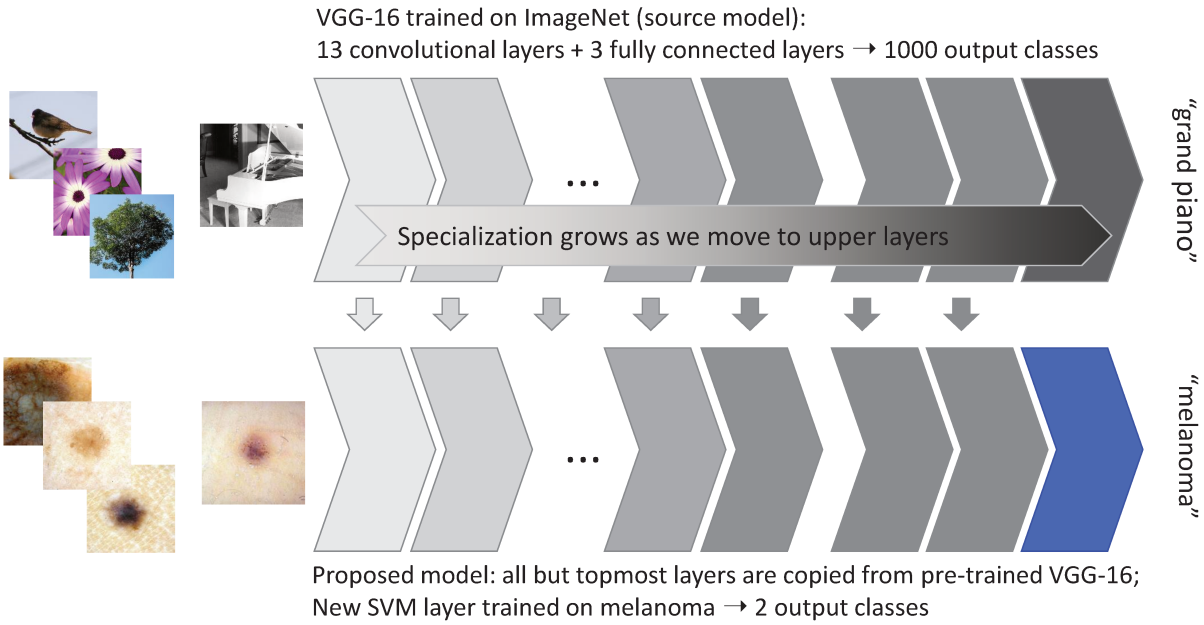


Figure 5 – A typical melanoma screening classifier using Deep Learning and Transfer Learning. We transfer knowledge from a Deep Neural Network pretrained on ImageNet (top row). All but the topmost layers are used to extract features that are fed to a new SVM layer, which is trained for the new task of melanoma screening. Image reproduced from Fornaciali et al. [2016].

by maximum likelihood, without need of biological justifications, and that is the viewpoint we adopt here. On the simplest DNNs, each neuron i on layer ℓ may be described by the function $f_{i,\ell}(x) : \mathbb{R}^{n_{\ell-1}} \to \mathbb{R} = \psi_\ell(W_{i,\ell} \cdot x_{\ell-1} + b_{i,\ell})$, where the input vector $x_{\ell-1}$ gathers the output of all neurons on the previous layer $\ell - 1$ (the input image, if $\ell = 1$), the vector $W_{i,\ell}$ and scalar bias $b_{i,\ell}$ are the weights of the neuron — the values that are effectively learned during training, and ψ_ℓ is a fixed (not learned) non-linear function, often a sigmoid or a ramp function. On the DNNs employed for image classification the description is a little more complex because the lower neurons are based on *convolutions*, but the fundamentals do not change. We refer the reader to LeCun et al. [2015] for a complete description. We show the common pipeline in 5.

The weights are learned by minimizing the error of the network predictions over the training set, a highly non-convex optimization that must be approached numerically, often with gradient descent methods. The gradient computations are made feasible by a technique called *backpropagation*, a clever way to use the chain rule for computing the derivative of the prediction error. The whole procedure, as it finds the weights that best adjust the model to the training set, is conceptually similar to a maximum likelihood estimation. As such, large models require huge training sets, and still are liable to under or overfitting. The craftsmanship of DNNs is delicate, full of ad-hoc "tricks of the trade" for regularizing and optimizing the models.

The greediness of DNNs for labelled data makes them cumbersome for medical applications, due to the costs of acquiring such data. As we mentioned, DNNs thrive with

hundreds of thousands, up to several millions learning samples, while a medical dataset of a thousand samples is considered large! That difficulty is, however, addressed by *transfer learning*, a technique that allows recycling knowledge from one task to others. In visual tasks, the low-level layers of a DNN tend to be fairly general; specialization increases as we move up in the networks Yosinski et al. [2014]. That suggests we may learn the general layers on *any* large dataset, and then learn only the specialized upper layers on the small medical dataset. A straightforward strategy for transfer learning is freezing the weights of a pretrained deep neural network up to a chosen layer, replacing and retraining the other layers for the new task. The new "layers" do not have to be themselves neural: other classifiers are perfectly acceptable.

More sophisticated techniques, which fine-tune the lower layers instead of freezing them, are available [Yosinski et al., 2014]. That means allowing those layers to evolve, expecting them to slightly adapt to the new task. That tends to correct any biases learned from the original task. When fine-tuning, it is easier to keep the entire network neural, to simplify the computation of the error gradients. Fine-tuning a DNN often leads to better results, which is also true for automated melanoma screening [Menegola et al., 2017a].

Besides fine-tune and transfer learning, other approaches enhanced the deep learning pipeline. One of the most relevant is data augmentation. Data augmentation are simple procedures — like translations, rotations, mirroring, and others — that generate new images from an input image. This technique allows the growth of databases without the difficulties inherent in new data acquisition. When using data augmentation in deep learning, the model *learns to become invariant* to image nuisances, replacing the traditional BoVW pre-processing steps, which *remove* the image nuisances.

2.3.2 Other Approaches

Before DNNs, other approaches dominated the literature of image classification. The most common are CBIR and the already introduced BoVW. Automated melanoma screening works employed those approaches for many years, CBIR being explored primarily in commercial applications. Their usages shape the related art as it currently is. Although classic, they are still being explored until now for melanoma screening (we refer the reader to the survey of Barata et al. [2018] for further information).

Sivic and Zisserman [2003] and Csurka et al. [2004] first proposed the BoVW model, appealing to an intuitive notion borrowed from textual information retrieval: the same way a text document can be usefully described as an unordered bag of words, an image could be described as a bag of *visual* words, without concern to geometric structure. That suggestive metaphor helped to popularize the model, although ultimately the parallels between words and visual words were weak. While words are intrinsically tied to semantics, visual words are the result of quantizing feature spaces based upon the

appearance of images, not their meaning [Avila et al., 2013].

Still, the model was very successful. The classical BoVW uses local detectors/descriptor to extract a large amount of discriminating features from small patches all around each image, and then applies an aggressive quantization on the space of those features to establish a visual vocabulary. The vocabulary allows finding visual words (*coding*), that are then aggregated into a frequency histogram (*pooling*). In a less colorful, but more rigorous view, the visual vocabulary is a codebook over the vector space where the low-level features reside, which allows establishing discrete codewords that can be counted. Figure 6 illustrates the pipeline.

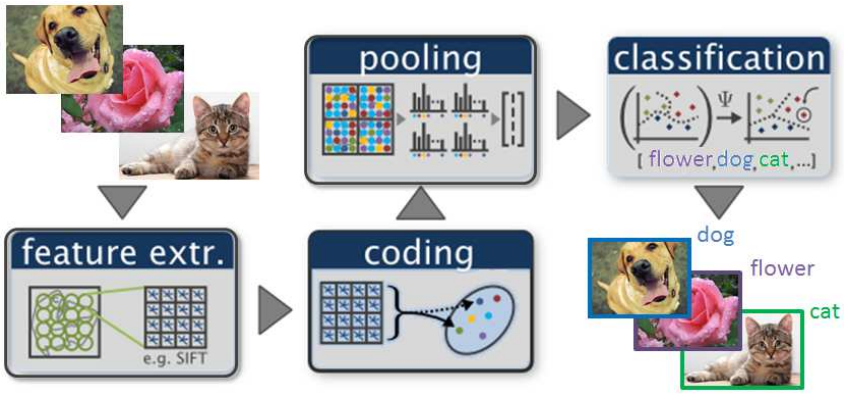


Figure 6 – Classical representation of a BoVW-based model. The feature extraction is the low-level stage. The mid-level is decomposed in the coding and pooling steps and the classification is generally done by a supervised method (e.g. SVM). Image reproduced from Fornaciali [2015].

Traditional BoVW ignores all large-scale geometric structure, gaining in generality and robustness but losing in discriminating power. Many extensions appeared, seeking to regain discriminating power without losing generality [Avila et al., 2013; Lazebnik et al., 2006; Perronnin et al., 2010; Jégou et al., 2010; Picard and Gosselin, 2011; Li et al., 2015].

Regarding CBIR techniques, the term "content-based image retrieval" seems to have originated in 1992 to describe experiments into automatic retrieval of images from a database [Eakins and Graham, 1999]. "Content-based" means that the search analyzes the contents of the image rather than the metadata such as keywords, tags, or textual descriptions associated with the image.

The general procedure (see Figure 7) is similar to the offline (training) / online (testing) steps of the typical image classification pipeline. CBIR also relies on feature extraction to describe the content of the images mathematically. Nevertheless, the aim is not to process such features to construct a predictive model, but to store the raw features in a pre-processed dataset. When new images are presented to the system, it extracts their features in the same way of the training phase and compares the similarity of the new features to those previously stored. Then, the system answers: instead of predicting a class for the querying image, the system returns other images most similar to the queried

one.

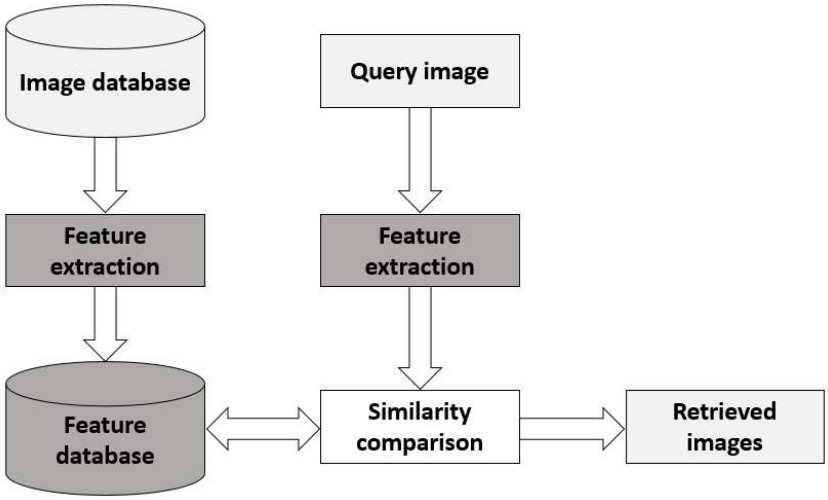


Figure 7 – Classical representation of a CBIR system. Image adapted from Kumar and Singh [2016].

The most common method for comparing two images is using an image distance measure. An image distance measure compares the similarity of two images in various dimensions such as color, texture, shape, and others. For example, a distance of 0 signifies an exact match with the query, concerning the considered dimensions; a value greater than 0 indicates various degrees of (minor) similarities between the images.

Most commercial solutions for automatic melanoma screening are based on CBIR. CBIR may be easier for doctors to accept, as its mechanism of operation is more natural to explain. Even the rational response is similar to the subjective process of physicians looking at a new lesion: they recall similar lesions/cases to provide a diagnosis of the lesion in question.

However, we must remember that the techniques of Deep Learning surpass the performance of CBIR; also, there are types of skin lesions that simulate other types, generating false positives. Still, we can not deny the success of commercial solutions — maybe exactly because they did not commit to the diagnosis, instead just counting on the human to make the final decision, they preserve the doctor's sense of agency.

2.4 History of Automated Melanoma Screening

This section describes the main facts, along with the history of automated melanoma screening. Here we describe the most relevant events, relating them, and showing how such facts contributed to shape and to understand the literature of automated melanoma screening as it is today. Figure 8 highlights the most relevant years.

The first literature description of using a computer to assist in the diagnosis of skin lesions was around **1985** [Day and Barbour, 2000; Korotkov and Garcia, 2012]. The idea

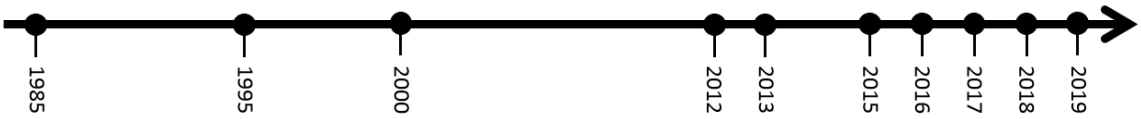


Figure 8 – The automated melanoma screening timeline, highlighting the years with the main facts.

was to feed a machine with skin lesion images. The computer captured features thought to be characteristic of malignancy, classifying the lesions and reporting the diagnosis. Those concepts matured over the years, especially around 1990, maybe due to more "availability" of skin lesion images. Several works emerged at the beginning of the '90s, mostly based on the ABCD Clinical Rule [Friedman et al., 1985] and Artificial Neural Networks. They all used small datasets of clinical images.

1995 was a remarkable year for melanoma screening, due to the emergence of dermoscopy, which enabled better visualization of the skin lesions [Day and Barbour, 2000]. From that point, the first dermoscopic images appeared, and started being adopted by automated models. Dermoscopy enhanced the accuracy of both human doctors and automated models [Sinz et al., 2017].

In **2000**, the first survey of automated melanoma screening was published by Day and Barbour [2000]. That review summarizes the literature from the '90s, alleging that although an automated tool to be used on the real-world seemed close to being delivered, the difficulties faced by the researchers would postpone that aim for more time. The authors argue that much of the work reported in the literature could have been spared if previous works had not taken shortcuts in describing their methods. With this, the literature seemed to advance slowly, since each work is reinventing the wheel. The main issues were: (a) lack of standardization on the testing sets, (b) lack of details describing proposed methods and (c) usage of small datasets to validate the models. As we are going to see, those issues are still open today, but new efforts are improving how the literature deal with them. That survey also reported an interesting phenomenon: until 1995 works used primarily clinical images; after 1995 works used dermoscopic images. That transition reflects a change of attitude in medical practice by doctors, as dermoscopy came to be accepted as the gold standard for diagnosis.

Jumping to **2012**, there was an important milestone for Computer Vision community: Krizhevsky et al. [2012] won the classification task of the *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC), employing a large, deep convolutional neural network, considerably improving the art at the time. Since then, this technique has been adopted in several Computer Vision tasks, but its usage on automated melanoma screening would be delayed for a couple of years.

After twelve years from the first survey, the literature of automated melanoma screening was revisited and summarized again by Korotkov and Garcia [2012]. Their work

surveyed the recent art, now based upon several hand-crafted feature descriptors designed to capture color, texture, symmetry, and other aspects that could mean malignancy of a lesion according to the medical literature. In the previous years, the Bag-of-Visual-Words models [Sivic and Zisserman, 2003; Csurka et al., 2004] emerged as the state-of-the-art approach to image classification, as well as melanoma screening. The authors conclude that the main obstacle to the more dynamic development of such field is the lack of a dataset for standardized evaluation of the proposed methods. Those datasets should be constructed based on the criteria of Malvehy et al. [2007], which include segmentation ground-truth, the final diagnostic and description of dermatoscopic features.

The lack of standardized data is, in fact, a severe problem for benchmarking. Fortunately, that issue is starting to disappear, since public datasets of skin cancer images, with structured and reliable information, are being constructed. The first public dataset to meet the criteria described earlier was released on **2013**: the PH2 Dataset [Mendonca et al., 2013]. Although it is a small dataset (200 dermoscopic images), it was enough to train and evaluate the models being developed until then. That justifies the number of published papers that reported their results on that dataset.

If the emergence of the PH2 database helped to converge the efforts of the literature, standardizing the results report, the year **2015** was deeply marked by the publication of the first articles using deep learning techniques for automatic melanoma screening [Masood et al., 2015; Codella et al., 2015]. The enthusiasm for deep learning has begun to bypass the approaches based on Bags-of-Visual-Words and to redeem the methods of Neural Networks, now in the version of Deep Neural Networks.

However, the new technique required many data to evaluate the approach. The old datasets were too small for the problem, causing difficulties to train the models or causing overfitting. Authors used non-standardized methods to artificially increase the datasets (data augmentation) or employed private datasets that were not shared (and which could not be quality assessed). Those facts unbalanced the literature efforts towards papers' comparability. Fortunately, the literature was graced in **2016** with the publication of the *International Skin Imaging Collaboration Program* (ISIC Archive) [isi]. The ISIC Archive is one of the biggest public datasets of skin lesions, containing more than 15,000 images. The institution was responsible for organizing the 1st edition of the *ISIC Challenge on Skin Lesion Analysis Towards Melanoma Detection* [Gutman et al., 2016], an open challenge of automatic classification of skin lesions, standardizing the conditions for the evaluation of competing methods. That was undoubtedly a unique achievement since such condition was needed for years to facilitate the evaluation of methods and allow the consistent advancement of literature. The ISIC Challenge editions have been repeated annually since then [Codella et al., 2018, 2019].

Since the first uses of deep learning for automated screening, much discussion

about the validity of such methods was raised, mainly due to the difficulty of explaining the technique's operation and justifying its way of working. In that approach, although the learned features mean to a machine, they may not make sense to humans. However, we could not ignore the significant performance gains. The improvements concerning the accuracy, AUCs, and other metrics reawakened the possibility of the machine being as good as or superior to humans trained to identify melanomas. Even with open public challenges to evaluate new works under the same conditions, the actual power of such approach was not clear yet. In **2017**, Esteva et at. Esteva et al. [2017] trained a deep neural network with more than 130,000 skin lesion images (mixing clinical and dermatoscopic ones), achieving an AUC>91%. They evaluated their method against 21 dermatologists, outperforming them. That impressive result was published in Nature, generating a series of medical publications discussing the benefits and implications of such statement.

Motivated by the work of Esteva et al. and the growing popularization of deep learning approaches, we proposed at the end of **2017** a robust design analysis of deep learning based solutions for automated melanoma screening [Valle et al., 2017]. We investigated the most relevant characteristics to boost the results of such techniques. We found that the depth of the deep architectures, coupled with the greater abundance of data, is crucial to the performance of the methods.

From 2017 onwards, deep learning solutions have definitively become a standard for the melanoma screening community. Other studies reporting higher machine performances than dermatologists also appeared in the literature in both **2018** [Haenssle et al., 2018] and **2019** [Tschandl et al., 2019]. Due to the difficulties of such a comparative study — and the amount of data required to achieve such performance — these were the only published works in this regard. It reveals that, despite the initiatives for data distribution, access to this type of material is still limited today. However, it is proven the viability of the automatic solution to the problem.

2.5 Recent Advances

In 2015, DNNs and other Deep Learning architectures started to appear as novel methods for automated melanoma screening [Masood et al., 2015; Codella et al., 2015; Premaladha and Ravichandran, 2016]. Two works ([Masood et al., 2015; Premaladha and Ravichandran, 2016]) employ DNNs still in the context of complex, traditional pipelines, based upon segmentation, preprocessing, etc. The work of Codella et al. [2015] marks, in that sense, a more definitive depart towards modern Computer Vision: a streamlined pipeline, with a modern BoVW on one hand, and a DNN+transfer learning on the other hand.

Existing works based on DNNs either train deep networks from scratch [Sabbaghi

et al., 2016; Nasr-Esfahani et al., 2016; Jia and Shen, 2017]; or reuse the weights from pre-trained networks [Codella et al., 2015; Yu et al., 2017; Esteva et al., 2017; Menegola et al., 2017a; Lopez et al., 2017; Harangi, 2017; Codella et al., 2017], in a scheme called **transfer learning**.

Transfer learning is usually preferred, as it alleviates the main issue of deep learning for melanoma screening: way too small datasets — most often comprising a few thousand samples. (Contrast that with the ImageNet dataset, employed to evaluate deep networks, with more than *a million* samples.) Training from scratch is preferable only when attempting new architectures, or when avoiding external data due to legal/scientific issues. Menegola et al. [2017a] explain and evaluate transfer learning for automated screening in more detail.

Whether using transfer or not, works vary widely in their choice of deep-learning architecture, from the relatively shallow (for today's standards) VGG [Yu et al., 2017; Menegola et al., 2017a; Ge et al., 2017a; Lopez et al., 2017], mid-range GoogLeNet [Yu et al., 2017; Esteva et al., 2017; Harangi, 2017; Yang et al., 2017; Vasconcelos and Vasconcelos, 2017], until the deeper ResNet [Harangi, 2017; Matsunaga et al., 2017; Menegola et al., 2017b; Bi et al., 2017; Codella et al., 2017] or Inception [Esteva et al., 2017; Menegola et al., 2017b; DeVries and Ramachandram, 2017]. On the one hand, more recent architectures tend to be deeper, and to yield better accuracies; on the other hand, they require more data and are more difficult to parameterize and train. Although high-level frameworks for deep learning have simplified training those networks, a good deal of craftsmanship is still involved. With the increasing availability of pre-trained networks, the choice of architecture becomes a complex and meticulous task. Networks with high performance in their origin task do not necessarily generate the most robust melanoma classifiers. [Perez et al., 2019] proposed a recent work evaluating that phenomenon.

Data augmentation is another technique used to bypass the need for data, while also enhancing networks' invariance properties. Augmentation creates a myriad of new samples by applying random distortions (e.g., rotations, crops, resizes, color changes) to the existing samples. Augmentation provides best performance when applied to both train and test samples, being more common on most recent melanoma screening works [Kawahara et al., 2016; Nasr-Esfahani et al., 2016; Menegola et al., 2017a,b; Bi et al., 2017; DeVries and Ramachandram, 2017]. Train-only augmentation is still very common [Yu et al., 2017; Esteva et al., 2017; Ge et al., 2017a; Lopez et al., 2017; Díaz, 2017; Yang et al., 2017; Vasconcelos and Vasconcelos, 2017]. Data augmentation is, indeed, a simple but effective approach to boost classifiers' performance. Some works propose new augmentations for automated melanoma screening [Perez et al., 2018; Vasconcelos and Vasconcelos, 2017].

As we have seen, Works based on global features or bags of visual words often pre-

process the images (some recent works using deep learning do it as well) to reduce noise, remove artifacts (e.g., hair), enhance brightness and color, or highlight structures [Abbas et al., 2013; Wighton et al., 2011; Sabbaghi et al., 2016; Matsunaga et al., 2017; Jia and Shen, 2017; Yoshida et al., 2016]. The deep-learning ethos usually forgoes that kind of "hand-made" preprocessing, relying instead on networks' abilities to learn those invariances — with the help of data augmentation if needed.

On the other hand, segmentation as preprocessing is common on deep-learning for automated screening [Nasr-Esfahani et al., 2016; Yang et al., 2017], sometimes employing a dedicated network to segment the lesion before forwarding it to the classification network [Yu et al., 2017; Díaz, 2017; Codella et al., 2017]. Those works usually report improved accuracies. We also evaluate the impact of **using segmentation** to help classification [Valle et al., 2017].

If ad-hoc preprocessing (e.g., hair removal) is atypical in deep-learning, *statistical* preprocessing is very common. Many networks fail to converge if the expected value of input data is too far from zero. Learning an average input vector during training set and subtracting it from each input is standard, and performing a comparable procedure for standard deviations is usual. The procedure is so routine, that with rare exceptions [Kawahara et al., 2016; Menegola et al., 2017b], authors do not even mention it.

Deep network architectures can directly provide the classification decisions, or can provide features for the final classifier — often SVM. Both the former [Esteva et al., 2017; Lopez et al., 2017; Harangi, 2017], and the latter [Sabbaghi et al., 2016; Ge et al., 2017a; Menegola et al., 2017a; Codella et al., 2017] procedures are readily found for melanoma screening. Also common, are **ensemble techniques**, which fuse the results from several classifiers into a final decision [Codella et al., 2015; Matsunaga et al., 2017; Menegola et al., 2017b; Bi et al., 2017; DeVries and Ramachandram, 2017; Harangi, 2017].

The evaluation of automated melanoma screening methods evolved a lot since the emergence of modern techniques. The datasets employed started to be — for the first time — somewhat standardized. Literature employing moder techniques uses mainly 3 datasets on the experiments: ISIC Challenge 2016 [Gutman et al., 2016], ISIC Challenge 2017 [Codella et al., 2018] or a subset of the ISIC Archive [isi]. Clearly, other datasets can be found, like some particular datasets of Hospitals. Three cases worth to be mentioned: on one hand Esteva et al. [2017] and Ge et al. [2017b], with datasets of more than 130,000 and 30,000 images respectively, unfortunately, both sets are private. On the other hand, Nass et al. used a small dataset of only 170 images (freely available), but that dataset size raise questions of overfitting due to small amount of data to enable effective learning with data greedy models.

Since the variation of datasets on literature is markedly lower, benchmarking for new works turned feasible, specially after the ISIC Challenge 2016, when authors started

to report the method performance on pre-stablished training and testing sets, enabling comparisons on regular basis. Such challenges also standardized the evaluation metrics: literature continued to report results regarding SE and SP, but AUC started to be systematically reported.

If reproducibility was an issue, a more mature literature is giving signals that this won't be a problem anymore: evaluation is being performed on same datasets and metrics, ISIC Challenge 2017 forced the competitors to publish reports detailing their methods and the usage of pre-trained networks enables a facilitated dialogue between researchers. However, since the models became more complex, the level of detailing must be higher in order to promote reproducibility. Fortunately, code sharing is being disseminated. Still, no work indicated concerns about developing reliable and less error prone code.

Although reproducibility issues have decreased, other problems emerged or became more evident, such as the lack of collaboration with poorly constructed medical teams and experimental designs. We also discuss such questions on Section 3.6.

Modern approaches should investigate new ways to incorporate medical knowledge in computerized tools. We know from experience that to base the feature extraction in medical rules is not relevant for deep learning, but the usage of medical information in other format can be useful. For example, Yoshida et al. [2016] proposed a simple preprocessing method especially designed for automated melanoma screening through deep learning: the alignment of the image with the major axis of the skin lesions. They argue that the proposed method improved the skin lesion classification AUC up to 5.8%, because that enables neural networks to easily extract symmetry data of the lesions, and melanomas tend to be asymmetrical.

Concluding, if the literature of melanoma screening became less skeptical about the power of an automated tool, they also became more demanding in terms of rigor in evaluation. Still, more work should be done in order to deliver an automated melanoma screening for the real-world. Other applications of such tools — as following-up the patient and evolution of lesions — should be explored as well.

2.6 Conclusion

The literature of automated melanoma screening is vast, with many hundreds of papers along almost thirty years. Several subtopics compose the literature, like medical algorithms, image pre-processing, lesion segmentation, change detection, and lesion classification. Regarding lesion classification, there are several advances in terms of new technical approaches, datasets for benchmarking, and expectations about the effectiveness of automated systems.

It is particularly interesting to note the medical view of automated systems during the time: in the early 1990s, the idea of a system capable of detecting melanomas by image analysis seemed unfeasible. With the advancing of the years and the popularization of the automatic tools, there was an increase in the concern about the quality of such systems. From the advent of deep learning, automatic classifiers had an unprecedented leap of accuracy. When melanoma screening works used such techniques — together with the greater abundance of data — we could observe the proposition of systems as effective as dermatologists trained to recognize the disease. This moment was a watershed, and the medical community started to see the automatic systems not as a promise, but a reality.

We, on the other way, plead that the automated tools surely can help physicians to derive better decisions, and also help people without easy access to specialists. However, there is plenty to be done to enable automated melanoma screening in real scenarios.

In this work, we aim to address technical and methodological gaps that prevent the advance of existing art towards robust melanoma screening for real-world. We also bring other practical contributions, sharing our codes to promote and increase reproducibility in the related community.

We list in Chapter 3 the main challenges and the main existing efforts to alleviate them, especially those related to data availability. In Chapter 4 we describe our experiments and analyze our results.

3 Critical Appraisal of Existing Art

In the previous Chapter, we introduced the foundations of automated melanoma screening: the technical, medical, and legal aspects. We also went through the literature, analyzing the main works from the computational point of view. In this chapter, we will complement that survey, by understanding how medical and legal aspects affect current art in the quest towards real-world deployment.

If in many areas Academia and Industry often operate apart, in Artificial Intelligence, they cooperate closely. In the case of healthcare, however, this cooperation involves a third party: Regulatory Agencies, which are responsible for enforcing laws and rules to minimize risks to the public. In this chapter we will explore the roles and relationships of those partners, as well as understand how they accelerate or delay the release of new solutions for actual use.

We start our appraisal in Section 3.1 relating the first interdisciplinary approaches of the literature. Then we move to the legal aspects, describing the regulation process of medical devices (Section 3.2), and specific topics of clinical evaluations (Section 3.3). We introduce a brief history of AI for healthcare in Section 3.4 and future directions for interdisciplinarity in Section 3.5. The central part of this Chapter is in Section 3.6, which discuss the open questions of the literature and the main initiatives to address them. We end this Chapter in Section 3.7, summarizing our analysis and findings.

3.1 First Steps Towards Interdisciplinarity

Although an appraisal that denied existing interdisciplinarity between Computing and Medical Sciences would be over-pessimistic, we are still far away from the ideal levels of cooperation between those two disciplines. Conducting interdisciplinary research is not necessarily trivial. When we started working with melanoma screening in 2013, we felt that, at the time, establishing an interdisciplinary cooperation was an uphill struggle. Once, however, we were able to secure that cooperation, our understanding of the medical/biological aspects of melanoma sharply improved, leading to a greatly enriched research. Today, we feel that there is an increased interest in interdisciplinary cooperation from all parties, although challenges still exist in establishing a common language, common research practices, and a common research agenda.

From the computing view-point, the first collaborations were based on computer methods mimicking medical rules to identify melanomas (notoriously the ABCD Rule and the 7-Points Checklist). As we saw, that procedure was not necessarily optimal, because

reported results were not robust and/or accurate enough to clinical usage [Rosado et al., 2003; Menzies, 1999; Rajpara et al., 2009].

When automated systems abandoned that approach and started to move out to more sophisticated techniques, attempting to exactly reproduce medical procedures in machine algorithms became a hindrance. Literature achieved results never seen before, equaling to trained dermatologists [Esteva et al., 2017; Mar et al., 2017; Safran et al., 2017; Marchetti et al., 2018].

The case is: humans and machines learn in different ways. So, should we leave Artificial Intelligence systems to learn predictive models on their own? The problem is that deep learning techniques learn from features that do not leave a track of how the decision was taken [1]. That raises serious questions if the machine is identifying skin cancers for the right biological reasons.

We know from experience that forcing the machines to act like a human being is not a good idea concerning the predictive power of the method. Nevertheless, we should incorporate medical knowledge in specific ways to boost automated results. Such approach could improve machine learning models [Yoshida et al., 2016]. A good hint would be teaching the machine the identify specific patterns that indicate when a lesion is a melanoma or not [Kharazmi et al., 2018], since machine learning models are suitable for that kind of task and Pattern Recognition is the most reliable technique to teach dermoscopy [Carli et al., 2003].

We should investigate the medical knowledge in a way not to reproduce it with machines, but to understand and use it wisely on automated approaches. Dermatologists and computer scientists should work together toward that goal. For example, in a context of referrability, epidemiological data must be incorporated even if out of the predictive model, in the way of a form that translates an implicit heuristic to calibrate the predictive model to better suit the patient or the case. Physicians already do that implicitly during anamnesis, trying to identify if a patient is more susceptible for developing the disease or not. Including such data into automated models is an important avenue for future works [Mar et al., 2017].

From the medical viewpoint, all recent surveys agree that modern techniques lead to powerful automated melanoma screening tools, usually being as accurate as or even better than dermatologists [Mar et al., 2017; Safran et al., 2017; Marchetti et al., 2018]. That acceptance has become even more common after the highly publicized paper of Esteva et al. [2017] on Nature. Nevertheless, automated melanoma screening for the real-world must be strictly evaluated, given the risks of misclassification and the lack of explicability of the current approaches. Those evaluations could be a fruitful research line [Safran et al.,

[1] https://www.technologyreview.com/s/604271/deep-learning-is-a-black-box-but-health-care-wont-mind/

2017].

3.2 Medical Device Regulation Process

For regulation purposes, a medical device is any instrument, apparatus, software, material or article used alone or in combination, including the software intended by its manufacturer to diagnose, prevent, monitoring, treatment or disease relieve among others [ISO 13485:2003]. The development and launching of medical devices rely on norms established by international standards and national regulatory agencies.

The **international standards** apply to the whole world. Consequently, any given region or country could adopt them, perhaps with modifications or limitations. The most famous standards are the The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC).

Regarding market share in the Western Hemisphere, the most relevant **regulatory agencies** are in Europe, USA, and Brazil[2]. They are different but retain several similarities, allowing their analysis under the same 4 steps-framework: (a) definition of the medical device intention of use and its characteristics, (b) classification based on its risks to the health professionals or the patients, (c) regulatory requirements, and (d) the regulatory process itself.

Here we do not aim to detail each process. We are going to illustrate the process by picking the American case[3], as it is the most straightforward one. Devesa [2014] lists the similarities and peculiarities of each process.

For melanoma screening, the medical device would be software, intended to be used as the diagnostic aid of skin cancer, through the analysis of skin lesion images. Since a misclassification would be dangerous to the patient, the risk offered by the system is high, indicating that it would be classified into a high-level class, regardless of the country regulating it.

Due to the risk to life potently offered by the system, the USA regulatory process must be fully followed. That includes a 510(k) Premarket Notification and a Premarket Approval (PMA). Each step aims to ensure that the proposed system is highly evaluated and has acceptable levels of quality. Nevertheless, the processes do not indicate the desired steps to assess the system. There are two types of evaluation: laboratory testing and human data evaluation. The first is something similar to what is done now: researchers evaluating their methods in controlled ways (experiments). The last relates to clinical trials, evaluating proposed methods on real life.

[2] To keep our scope feasible, we will not analyze Asian markets in this text.
[3] Regularized by FDA: https://www.fda.gov/

For the 510(k) clearance [4], human data is optional, but PMA requires it. So, literature is typically prepared for a 510(k) clearance, evaluating new methods under controlled approaches: using provided datasets, constructing their experimental design, comparing results. More sophisticated evaluations include comparisons of systems to dermatologists. Nevertheless, the "gold standard" clinical trials of automated melanoma screening were never seen, so no system proposed until now is ready for a PMA.

The USA process is rigorous and liberal at the same time: it is rigorous in the sense that requires clear quality and effectiveness evidence, but is liberal in the sense of not describing *how* to deliver it (that is, how should we evaluate CAD systems) [Devesa, 2014].

In practical terms, the manufacturer must prove to the government that the system was validated and the results are reliable enough to demonstrate its effectiveness. In that sense, works that report their findings on rational experimental design and count with expert validation (dermatologist comparisons) [Esteva et al., 2017; Marchetti et al., 2018], would have their quality and effectiveness more easily proven.

Regarding medical software development, some global standards must be followed. The most famous are ISO 9001:2008; ISO 13485:2003; IEC 62304:2006. Such standards indicate maintenance requirements, risk management, problem tracking, and problem-solving that are not even followed by academic studies of melanoma screening (at least, no evidence was found supporting that concerns on academia).

What do we learn from this context? From legal regulation, there is a concern to prove that the automatic screening system works appropriately, but there is no guideline to follow. Given such freedom, comparison with previous works may be an alternative. The market cares about the quality of system development (aspects of the code itself), which is generally not addressed in academia.

3.3 Clinical Trials

Clinical evaluation is the most advanced approach to validate a medical device regarding its effectiveness and reliability. It is recommended, according to the medical devices' usage critically, so higher the critically, the more compulsory is the necessity for a clinical evaluation. Clinical validation is necessary for any SaMD.

The process follows three steps, as illustrated in Figure 9. Those steps are not performed during regulation time. At the regulation time, all the procedures to develop and evaluate the system are investigated according to the documentation provided by the

[4] A 510(k) clearance is a simple process of premarketing a medical device that does not need premarket approval. The 510(k) process aims to demonstrate that the device to be marketed is at least as safe and effective, that is, substantially equivalent, to a legally marketed device.

manufacturer, which describes the whole process in a textual and/or illustrated way.

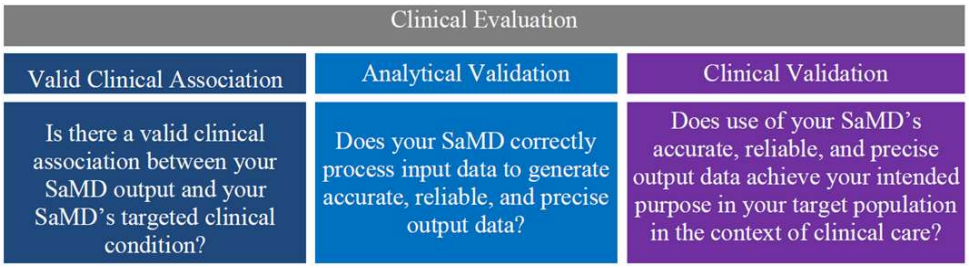


Figure 9 – The clinical evaluation process.

First, the manufacturer needs to demonstrate a valid clinical association of the new system's output to the target condition. The analytical validation aims to access the development process. Finally, clinical validation shows how the system was validated in real-case scenarios.

In other words, the first step is to show that the final system makes sense. That is, to access if the intended usage of the system aligns to the context of the health condition to which the system is directed. The second step is a typical validation and verification process of software development: the assessment aims to determine if the software was correctly constructed, that is, if the development followed a reasonable process demonstrating that the software meets its specifications and that specifications conform to user needs and intended uses.

The third and last step is clinical validation. It measures the ability of a SaMD to yield a clinically meaningful output regarding the target condition and the software intended usages. Either can demonstrate clinical validation:

- Referencing existing data from studies conducted for the same intended use;
- Referencing existing data from studies conducted for a different intended use, where extrapolation of such data can be justified;
- Generating new clinical data for a specific intended use.

Although the typical process has the above three steps, sometimes the clinical evaluation has a fourth step: the independent review. According to the FDA Guidance for Clinical Evaluation of SaMD, *the recommendation for independent review highlights where the evidence generated from the clinical evaluation of the SaMD should be reviewed by someone who has not been significantly involved in the development of the SaMD, and who does not have anything to gain from the SaMD, and who can objectively assess the SaMD's intended purpose and the conformity with the overall clinical evaluation evidence.* Also, the *independent review is more important for SaMD that "Treats/Diagnoses Serious*

and Critical" health care situations and conditions and SaMD that "Drives Critical" health care situations and conditions.

Figure 9 summarizes the situations when an independent review of the clinical evaluation is more recommended. Undoubtedly, automated skin lesion analysis for skin cancer triage is a situation that requires a clinical evaluation and its independent review.

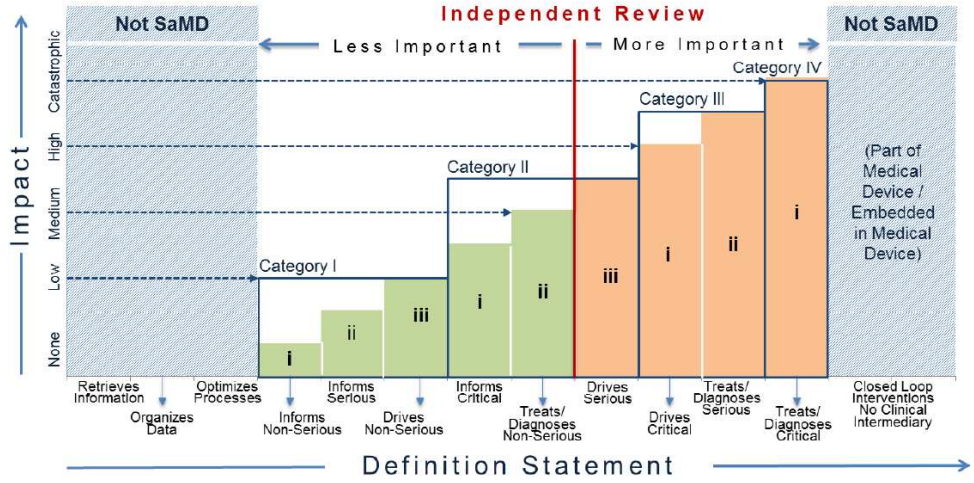


Figure 10 – Risk Based Approach to Importance of Independent Review.

3.4 History of AI for Healthcare

At the beginning of the technological introduction in medicine, electronic hardware responded for most of the innovation in medical devices. As technology progressed, software became progressively more important, and its role in control/management/final application became more evident, thus requiring specific regulations. The year of **2006** was a turning point, with the creation of IEC 62.304/2006 [IEC 62304:2006], regulating the process of development of medical software — either intended for standalone use, or in conjunction with hardware.

The following years saw a burst in the number of software-only medical devices, mainly due to the proliferation of mobile apps. In response, in **2011**, the new category *Software as a Medical Device* (SaMD) was created [of Health and Services], contemplating aids to diagnosis, screening, monitoring, determination of predisposition, prognosis, prediction and determination of physiological status.

The concept of SaMD was consolidated in **2013** with the creation of International Medical Device Regulators Forum (IMDRF) [5], led by the FDA. The IMDRF establishes the key definition points of a SaMD, the risk classifications, the development quality management system, and clinical evaluations.

[5] International Medical Device Regulators Forum: http://www.imdrf.org/

Until then, a SaMD should follow the same process of regulation of an ordinary medical device, that is, it is expected to be fully finalized and extensively tested before initiating the regulation request.

In **2018** the FDA launched the *Breakthrough Device Designation*[6]: an iterative process that the producer can gain access to the FDA before official marketing approval submissions, to accelerate the development by removing doubts and enhancing the creation of new technologies.

Medical devices are eligible for Breakthrough Device Designation if **both** of the following criteria are met[7]:

1. Firstly, the device provides for more effective treatment or diagnosis of life-threatening or irreversibly debilitating human disease or conditions;

2. Secondly, the device also meets *at least one* of the following: (a) represents breakthrough technology; (b) no approved or cleared alternatives exist; (c) offers significant advantages over existing approved or cleared alternatives; (d) device availability is in the best interest of patients.

The shortcut is particularly relevant to SaMD, as the regulatory process approaches the incremental/agile process of software development.

Also in **2018**, the FDA approved the first fully AI-based device for healthcare: a system that detects diabetic retinopathy by eye images analyses and decides whether a patient should promptly visit an ophthalmologist or if they must repeat the exam in 12 months [Food and Administration, 2018]. This fact was a milestone in history and paved the way for the legalization of other systems.

3.5 Future Directions for Interdisciplinarity

Here we reach the core of the discussion: scientific literature in melanoma screening does not follow the robustness required by regulatory standards for medical devices. Although some works count with medical cooperation, the testing procedures are still limited to lab examination and stand-alone evaluation.

We need to highlight, however, that not all published work necessarily aims to develop a screening system for commercial use. However, since the commercial evaluation is an open process, why not to base it on state-of-the-art approaches found in literature?

[6] Breakthrough Device Designation: https://www.fda.gov/medical-devices/how-study-and-market-your-device/breakthrough-devices-program

[7] Breakthrough Devices Program: Is my device eligible? — https://www.fda.gov/medical-devices/how-study-and-market-your-device/breakthrough-devices-program#s3

So, if the literature is populated with non-standardized evaluation and/or with critical shortcuts, a new product quality evaluation is impaired in a vicious cycle, preventing the development of automated tools for real-world.

On the other hand, the evaluation of market solutions comprises bench tests and clinical trials. Bench tests, if not conducted properly, may carry the same problems of the academia, which we list on the next Section.

Clinical trials are the best approach to avoid hidden issues with experimental design and data usage. However, clinical trials must rely on statistical data and previous experiments, generally collected in the literature. Such reference may introduce noise in the perception of the real quality of the final systems, besides perpetuating vices of the literature that do not contribute to the real advance of the problem.

Although such relationship may seem obscure, the FDA process for regulation of software as a medical device subtly reveals it (Figure 11).

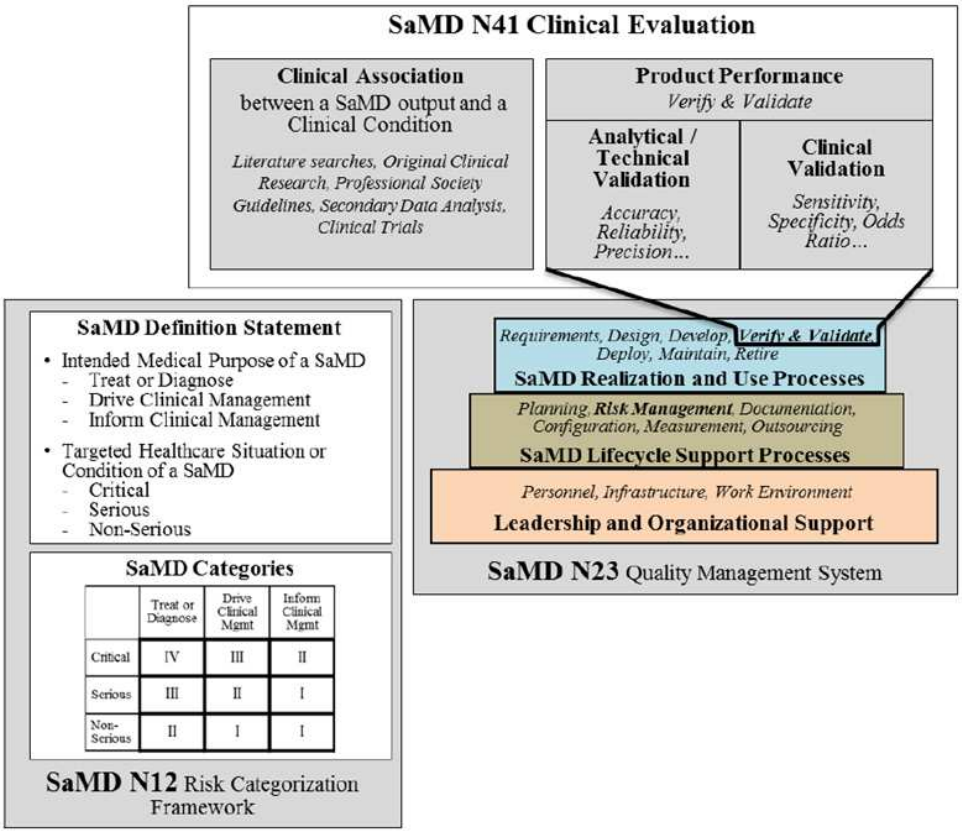


Figure 11 – The Software as a Medical Device regulation process.

At this point, the interaction and dependency between market and academia are clear: new clinical trials are new benchmarks that compose the literature, and the literature provides baseline results for comparing new clinical evaluation outcomes. The main issue with that interaction is the problems that the literature presents. Consulting, basing, or comparing with literature can perpetuate the vices and problems of existing art.

We should pay particular attention to the evaluation metric choices. There is not a consensus on literature about which are the best choices to evaluate automated skin lesion analyzers. While the medical opinion is extremely relevant, some metrics are not correctly used on tuning computer vision models, like the pair sensitivity and specificity.

We believe that such relationships will continue in the following years. The goal is, therefore, to eliminate the underlying problems.

In the next Section, we introduce the main gaps of the literature and the ongoing efforts to eliminate them (if any). Future work should address such gaps to push further the development of new solutions for automated melanoma screening.

3.6 Open Questions

In this Section, we list and discuss the open questions of current art. We divided them into two sets: the traditional challenges and the new challenges that emerged with the technology development. For each issue, we describe *why* it is a problem and *what* are the recent efforts to overcome it (if any).

3.6.1 Traditional Challenges

We start with classical issues, which have been present in the field since its inception. For most of those challenges, important initiatives have appeared to solve, or at least alleviate the problem.

1. **Experimental validity**: This is a severe issue because it is less evident than the others. Despite the existence of dozens of published papers about automated melanoma screening, the accumulated knowledge is not reliable enough to make decisions in the level of rigor needed for actual clinical practice, e.g., the level of rigor expected by regulatory agencies . Authors choose different metrics: most only report global accuracy — which is misleading when the positive and negative classes are highly unbalanced. If a dataset has 20 images of melanoma and 80 images of non-melanoma — a disproportion quite typical of melanoma datasets — one can classify all images as non-melanoma and still get 80% accuracy. Therefore, some authors prefer the pair Sensitivity–Specificity, which together, reveals such problems (in the example above, one would get 100% specificity, but 0% sensitivity). Some authors — us included — prefer to report the area under the ROC curve (**AUC**), which averages all possible Sensitivity–Specificity pairs given by the classifier, and measures, so to speak, the amount of "knowledge" the classifier has about the problem, without regard to its preference for sensitivity or specificity. More exotic metrics,

like the F-measure of information retrieval, also appear. Agreeing on metrics is a necessary, not sufficient, condition to render the works comparable.

Works exist which are still validated in datasets too small to provide reliable evidence. Further datasets (of any size) are often proprietary and not available for cross-examination. Comparisons would still be possible, in principle, if each research group ran previous techniques in their data, but authors will seldom share code (as will be discussed on the following topics).

2. **Lack of interaction between researchers and physicians**: This is by far the most severe issue that prevents the development of an automated tool for real-world usage. An automated melanoma classifier is not just an accurate, specialized image classification system: it is a software that handles with sensitive information; an implementation of procedures and methods; a new working tool for health professionals. How can it be successfully designed without the cooperation of all stakeholders?

   Technical personal must understand screening procedures; health professionals should be acquainted with the system usage; dermatologists should collaborate to train the system with relevant information in order to be accurate; the system itself should be carefully audited before actual use.

   Today, more interaction exists between those worlds, but it is not enough to deliver usable approaches. Current works usually compare their results with the performance of trained dermatologists to identify malignant lesions. The interaction is restricted to the final stages of the development process when it should occur since the beginning.

   Against the lack of interdisciplinary collaborations, an international effort deserves our attention: The International Skin Imaging Collaboration — ISIC Melanoma Project. They are an Academia–Industry partnership, involving both Medical and Computer Science, designed to facilitate the application of digital skin imaging to help reduce melanoma mortality. ISIC is developing standards to address the technologies, techniques, and terminology used in skin imaging with special attention to the issues of privacy and interoperability.

   ISIC also organize annual workshops in conferences focused in computer vision, aiming to promote cutting-edge researches in skin lesion analysis, including several toppics like dataset issues, modern image types and new algorithms. Those workshops also feature several prominent names in this research field, showcasing research trends and encouraging colaborations.

3. **Reproducibility**: Since most authors do not share code and/or data, attempts at comparison must incur in the expensive effort of *reimplementing* previous works.

That is not exclusive to melanoma screening but happens on all computing literature [Peng, 2011; Sandve et al., 2013], and even scientific literature in general [Allison et al., 2016]. However, even that is often impossible. The vast majority of works do not describe the details that allow reimplementation. That fact was already observed by Masood and Ali Al-Jumaily [2013], in works that spanned from 1993 to 2012; our exam of works from 2008 to 2016 found a situation just as dire [Fornaciali et al., 2016].

Due to the current complexity of the computer systems employed in scientific experiments, the reproducible paper without original code is a unicorn, a mythical creature. Even the most detailed description leaves out essential information, to respect page limits, or to avoid tedious/confusing the reader with a myriad of minutiae. Still, a single wrong parameter can collapse an entire experimental cathedral.

Examples of attitudes that disturb reproducibility are: missing details, results much below the ones reported originally, and failure to earn the cooperation of the author.

Reproducibility is an open issue until now, but it tends to soften due to initiatives such the ISIC Challenge, that forces the competitors to use the same dataset, the same metrics, the same testing set and also, started to require in 2017 a technical report about each submitted solution. The reports are not peer-reviewed, and there is not a guarantee that the methods are fully described, but the initiative is already a significant step towards reproducibility. Also, in the last years, the number of papers sharing code in public repositories increased.

4. **Lack of standardized public databases**: Data is the main issue towards a fair comparison between different approaches. If two works use different datasets to evaluate their proposals, benchmarking is unfeasible, and conclusions are limited. Even when the same dataset is employed, comparability can be severely hindered due to each work selecting unspecified subsets of the dataset. Some diagnostic classes (melanosis, recurrent nevus), or lesions with specific clinical characteristics (e.g., lesions on the palm of the hands, sole of the feet, or the genital mucosa) are notoriously problematic, even for humans. Hair, rulers, bandages or other artifacts may hinder the analysis — thus images containing those artifacts are often excluded. The issue is not the selection itself, but the uncontrolled variability brought by an *unknown* selection.

   Fortunately, this issue is the one closer to be eliminated. Since the emergence of public datasets, works tend to evaluate their approaches on a more regularized basis. In this sense, global initiatives (like the ISIC Archive), are precious, since they also provide standardization of testing sets and metrics.

5. **Mimicry of medical rules**: Most of the existing art often attempts to directly

mimic medical diagnostic rules, especially the ABCD Rule of Dermoscopy [Friedman et al., 1985], and the 7-Points Checklist [Argenziano et al., 1998]. Attempting to mimic human procedures leads to rigid processing pipelines based on segmentation → feature extraction → classification, which no longer reflects the state-of-the-art in Computer Vision. That is compounded by the use of features imitating medical reasoning (color, texture), which are known to underperform when compared to new options. Such approach was the first attempt to solve the problem since it is highly intuitive: we take a medical algorithm and try to implement it in a system that aims to reproduce step by step all actions needed to identify if a lesion is malignant or not.

Even if literature were successful in imitating the ABCD Rule, or the 7-Points Checklist, one could still argue that those rules were supplanted by the newer 3-Point Checklist [Soyer et al., 2004], and the Revisited 7-Point Checklist [Argenziano et al., 2011]. More telling is the fact that none of those rules "caught on" among doctors: traditional pattern analysis [Pehamberger et al., 1987] is still considered the most reliable technique to teach dermoscopy Carli et al. [2003].

Implementing medical rules as computer algorithms has the advantage of being explainable for a broad audience. If it is a question of explainability, new works of explainable deep learning can be an alternative path for future research. The incorporation of medical knowledge, however, should not be done only in terms of image processing, but heuristics that assist dermatologists in decision making. For example, what clinical data are relevant, and how the relationship between them may indicate the malignancy of a lesion? Again, such integration would only occur in a more cooperative scenario of researchers and physicians.

3.6.2 New Challenges

With the advancement of AI solutions for healthcare, new issues arise for the literature. The main points of attention are:

1. **Typical AI problems**: machine learning relies on data, and data carries inherent problems. Among them, the main thing is the quality of the data: how were they collected? How were they cataloged? Can we rely on the annotated diagnoses? Also, we know that every dataset has biases. What are the biases of medical data, and in particular, images of skin lesions?

   We can not fail to mention the concerns about the design of the final systems: how will they influence the doctor's decision? How will they influence the patient's treatment? Will that systems really improve the long-term health of a population?

2. **Healthcare providers resistance to AI adoption**: the opinion of physicians about automated systems is essential for the development of melanoma literature. Their concerns about the impacts of technology introduction are valid. While some authors, or especially the media, tend to disclose the advances of technology as "steps to replace doctors", we believe that this is not feasible. On the contrary, we believe that computational methods will be increasingly perfected to aid the routine of doctors and patients.

   Massive access to screening technology, however, deserves attention. The risk of the public using this technology in place of seeing a dermatologist exists. It can be combated by intensifying education actions for the population about the problem and importance of the visit to the doctor whenever a lesion raises suspicion. After all, the machine may even detect a malignant lesion, but will not treat it.

3. **Problems in validation protocols**: as technology becomes more mature, its use in real environments becomes more feasible. We have seen in this Chapter the process of regulation of medical devices and medical software. There is a veiled relationship between literature and regulation processes. If literature carries problems (and it does), the legal evaluation of new solutions can be hampered.

   To mitigate possible negative impacts of using technology, it is recommended to work on validations in two moments: 1) understand *what* AI recommended (evaluation metrics; compare with experts to see if it makes sense); 2) understand *why* AI recommended (explanations of how models work; combine AI and humans to get better decisions than either alone).

3.7 Conclusion

In this Chapter, we introduced the interdisciplinary aspects of the research for automated melanoma screening. We saw that literature deals timidly with this subject, with occasional interactions, which, while important, do not meet the needs for the development of robust solutions.

We also evaluated the relationship between the development of academic solutions and commercial products for melanoma screening. At first sight, these relations seem subtle, but the regulatory process, required for validation and commercialization of professional solutions, forces the interaction between the two worlds.

If we do not worry about the quality of the solutions developed, both professionally and academically, we will be contaminating the literature with failures that increasingly delay the adoption of solutions based entirely on Artificial Intelligence.

The current literature already presents problems that need to be overcome. Not surprisingly, the existing problems are also interdisciplinary.

In the next Chapter, we will describe our contributions to alleviate some of the problems identified above. Due to the nature of our research group, we focused mainly on the computational issues.

4 Advancing Machine Learning Models

In the early 2010s, the limited predictive power of computational models imposed the main barrier to the success of automated melanoma screening. Recent advances of the machine learning models removed such barriers, but we still face the challenge of scarcity of training data, in addition to methodological shortcomings in existing literature, such as lack of reproducibility.

From 2015 onwards, melanoma detection moved onto deep learning models, associated with transfer learning. Besides demanding vast amounts of data and computation, deep learning poses some challenges per se, like the regularization of thousands, or millions of parameters, and also the choice of several aspects of the architecture design. On the other hand, transfer learning also brings its challenges, like the choice of the source database, the layers to be transferred, and the fine adjustments to the target task.

This Chapter describes our efforts towards further improving machine learning models. Our main contribution is on the methodological analysis of current deep learning approaches. We identify the main gaps, proposing interventions that systematically improves classification results, even without profound changes in existing architectures. We argue that our contributions reduce the distances that prevent the use of current solutions in real environments. We also deliver new methodological questions that must be taken into account in the design of automatic medical screening systems.

We started working with automated melanoma screening in 2013. We concentrate our results from the first three years of research in Section 4.1, and we provide an overview of our recent works (2017 onwards) in Section 4.2.

To facilitate the reading of recent works, we concentrate the description of the datasets, computational resources, and basic experimental framework in Section 4.3. The following sections describe each hypothesis and how we address them. We end this chapter summarizing our contributions on Section 4.9.

4.1 Previous Works

Our research group has worked with automated melanoma screening since 2013. At the time, image classification was taking the first steps towards deep learning approaches. However, for specific domains — like medical tasks — BoVW-based models still emerged as the most promising solutions.

BoVW models, as we saw in Section 2.3, have limitations in learning the spatial distribution of visual information, and often require additional processing steps to

eliminate the images nuisances, such as hair and medical artifacts.

In this context, we explored modern extensions of BoVW models, proposing a simpler and more straightforward image analysis process, eliminating the need for complex preprocessing. We have achieved competitive results with current literature, marking our first publication in the community ([Fornaciali et al., 2014]). The maturation of the work culminated in the master's thesis of the present author ([Fornaciali, 2015]).

Still in 2015, we also released our first results using deep learning with transfer learning for automated melanoma screening [Carvalho, 2015].

From 2016 onwards, the automated melanoma community moved from traditional techniques towards deep learning, following the general trend of computer vision. Even with the profound change in the use of computational techniques, the literature of the time carried the vices of many years of research, accumulating difficulties of access to data and reproducibility. We analyze the literature of the time, discuss such problems, and propose directions to mitigate them [Fornaciali et al., 2016].

4.2 Recent Works: An Overview

From 2016 onwards, we completely adopted deep learning in our group. We started our journey investigating the impact of transfer learning on final results. Although much of the best art on automated melanoma screening employs some form of transfer learning, a systematic evaluation was missing. We detail our approach and main findings on Section 4.4.

In 2017 we participated in the 2nd Edition of the "ISIC Challenge on Skin Lesion Analysis Towards Melanoma Detection". Based on previous results, our proposal lies in a deep learning model with transferring learning from ImageNet. It relies on four pillars: data, models, tricks for improving the models' accuracy, and ensembling of partial results. Our approach took us to the first place on the "melanoma *Vs.* all" subtask, which is the most relevant one, regarding screening purposes. We detail our participation and main results on Section 4.5.

Our participation in the 2017 ISBI Challenge raised some questions of what are the main aspects that contribute to improving classification rates of methods based on deep learning. We investigate the methodological issues for designing and evaluating deep learning models for melanoma screening, by exploring nine choices often faced to design deep networks: model architecture, training dataset, image resolution, type of data augmentation, input normalization, use of segmentation, duration of training, additional use of SVM, and test data augmentation. Such analysis is the main methodological contribution of this work to the related art since we demonstrated with reliable experiments how

a successful melanoma classifier should be designed to increase robustness for real usage scenario. We detail our protocols and main findings on Section 4.6.

With the popularization of public databases of skin lesions and the results of the 2017 Challenge as an increasingly publicized baseline, new approaches emerged. In 2018, the ISIC Challenge launched a re-reading of the main problem: instead of detecting melanomas, the task became the identification of multiple skin lesions, malignant or benign. We argue that the existing methods are not ready yet for such a task since it is even harder than the original one. Nevertheless, we embraced the task and also participated in the Challenge, which we discuss in Section 4.7.

As expected, the performance of the competitors was not robust enough to prove the viability of the new task, even in controlled environments. One of the critical issues was, again, the availability of annotated data. Considering that the public datasets are limited, we questioned their overall quality, their representativeness of actual conditions of occurrence of skin diseases in the population, and what characteristics of such images the computational models are learning to recognize. It motivated experiments to illustrate such phenomena, which we have listed in section 4.8. The main results indicate that the data used in current researches have biases that drive computational models to recognize unintended patterns, making it challenging to generalize learning for use in real environments. We present such questions to the literature and hope they will guide future research.

4.3 Materials & Methods

Along this work, we collected several **data sources** to compose our own dataset. We employed 10 different data sources, which we detailed below. Altogether, we end up with approximately 30 thousand images, comprising almost 3 thousand melanomas.

- **EDRA Interactive Atlas of Dermoscopy** (Atlas): The Interactive Atlas of Dermoscopy [Argenziano et al., 2002] [1] is a multimedia guide (Booklet + CD-ROM) intended for training medical personnel to diagnose skin lesions. It has 1000+ clinical cases, each with at least two images of the lesion: close-up clinical image (acquired with a Nikon F3 camera mounted on a Wild M650 stereomicroscope), and dermoscopic image (acquired with a Dermaphot/Optotechnik dermoscope). Most images are 768 pixels wide × 512 high. Each case has clinical data, histopathological results, diagnosis, and level of difficulty. The latter measures how difficult (low, medium and high) the case is considered to diagnose by a trained human. The diagnoses include, besides melanoma (several subtypes), basal cell carcinoma, blue

[1] EDRA Interactive Atlas of Dermoscopy: `http://derm.cs.sfu.ca/Welcome.html`

nevus, Clark's nevus, combined nevus, congenital nevus, dermal nevus, dermatofibroma, lentigo, melanosis, recurrent nevus, Reed nevus, seborrheic keratosis, and vascular lesion.

- **PH2 Dataset**: The PH2 Dataset [Mendonca et al., 2013] [2] has 200 dermoscopic images (80 common nevi, 80 atypical nevi, and 40 melanomas), acquired at the Dermatology Service of Hospital Pedro Hispano/Portugal. The dataset also provides ground truths for segmenting the lesions.

- **Dermofit Image Library** (Dermofit): The Dermofit Image Library [Ballerini et al., 2013] [3] is a collection of 1,300 skin lesion images and their segmentation masks divided among 10 classes (Actinic Keratosis, Basal Cell Carcinoma, Melanocytic Nevus, Seborrhoeic Keratosis, Squamous Cell Carcinoma, Intraepithelial Carcinoma, Pyogenic Granuloma, Haemangioma, Dermatofibroma and Malignant Melanoma). The diagnoses were provided by expert dermatologists and dermatopathologists, generating a gold standard groundtruth.

- **IRMA Skin Lesion Dataset** (IRMA): this dataset was created by the Department of Medical Informatics, RWTH Aachen University. It has 747 dermoscopic images (being 187 melanomas). We employed this dataset in our preliminary experiments. Today it is not available. So, in order to promote reproducibility, we exclude this dataset in our recent works. Nevertheless, due to its usefulness, especially in times of low availability of data, we list its contribution in this work.

- **Kaggle Challenge for Diabetic Retinopathy Detection** (Retinopathy) dataset [4]: this Retinopathy dataset has a training set of 35,000+ high-resolution retina images taken under varying conditions. It is a benchmark for validation of new retinopathy detection systems. We only used this dataset to compare different image sources in transfer learning approaches for melanoma screening.

- **ISIC Project**: The "International Skin Imaging Collaboration: Melanoma Project (ISIC)" [5] is an academia/industry partnership, coordinated by the International Society for Digital Imaging of the Skin, to acquire and annotate skin lesion images. As of March 2016, $\sim$ 3000 images were available; today, more than $23,000$ images compose the dataset, collected from different leading clinical centers internationally, using a variety of devices for acquisition. Since 2016, this dataset is also increasing in the amount of information available for each lesion: segmentation masks and

[2] PH2 Dataset: `www.fc.up.pt/addi/ph2%20database.html`
[3] Dermofit Image Library: `https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html`
[4] https://www.kaggle.com/c/diabetic-retinopathy-detection/data
[5] ISIC Project: `www.isdis.net/isic-project`

maps over five dermoscopic attributes (pigment network, negative network, streaks, milia-like cysts, and globules) are available for smaller subsets of the dataset.

- **ISBI Challenge 2016 Part 3: Disease Classification** [Gutman et al., 2016]: it is a subset of 1,279 dermoscopy images from the *ISIC Project*. The Challenge Dataset contains 900 images for training (273 being melanomas) and 379 for testing (115 being melanomas).

- **ISIC 2017 Challenge** Codella et al. [2018]: the official 2017 challenge dataset, with 2,000 dermoscopic images (374 melanomas, 254 seborrheic keratoses, and 1,372 benign nevi).

- **ISIC 2018 Challenge** [Codella et al., 2019; Tschandl et al., 2018]: the official 2018 challenge dataset, with 10,015 dermoscopic images with 7 ground truth classification labels (1,113 melanomas, 1,099 benign keratosis, 6,705 benign nevi, 327 actinic keratosis, 514 basal cell carcinomas, 115 dermatofibromas and 142 vascular lesions).

- **Other Sources**: the ISIC 2018 Challenge dataset is extreme imbalanced, so we decided to gather extra images for the severely underrepresented classes (the last four of them). We found images browsing sources on the web, and asking for contributions from partner researchers in Medical Science. The web sources were **Dermatology Atlas** (www.atlasdermatologico.com.br), **Derm101** (www.derm101.com), **DermIS** (www.dermis.net/dermisroot). With that extra effort, we acquired additional 631 images, being 414 basal cell carcinomas, 26 actinic keratosis, 132 dermatofibromas and 59 vascular lesions. Although the final dataset continued seriously unbalanced, the proportion of underrepresented classes grew considerably.

Regarding the **computational resources**, the employed infrastructure varied along with the research, depending on resources available and external grants. Altogether, we used NVIDIA GPUs available at RECOD Lab: two Titan X Pascal, six Titan Xp, one Tesla K40, and for Tesla P100. We also used the NC6 (Tesla K80) and ND6 (Tesla P40) virtual machines provided by the Microsoft Azure Cloud platform. Part of the experiments related to our participation in the ISIC Challenge 2017 was performed at the LIP6/UPMC/Paris, which hosted Prof. Valle during most of the competition, and generously offered part of the needed resources.

Finally, our **methods** obeyed the traditional framework of machine learning experimental design: we have gathered the data currently available, separated into independent training, validation, and test sets. We train the models in the training set and follow their maturation in the validation set. In the end, we report the official results in the test set. We describe the particularities of each protocol in the following sections, experiment by experiment. We present the same structure in every Section: we introduce the motivation

and main objectives of the experiments, following the experimental proposal, concluding with the main findings and its analysis regarding the current art.

4.4 If knowledge is needed, which knowledge should be *transferred*?

The first works employing deep learning for automated melanoma screening were generally based on transfer learning, since the available datasets were small to train complex architectures from scratch. Moving from classical techniques to deep learning approaches introduced new baselines for modern art of melanoma screening, similar to what had been happening to other classification tasks since 2012.

What was unclear, however, was whether the improvement of the skin lesion classification results was due solely to the better features produced by the deep models or whether the transfer of knowledge from large databases carried the more significant part of the predictive power of the new approaches.

In that scenario, we investigated the impact of transfer learning on the final results. We investigated the presence of transfer, from which task the transfer is sourced, and the application of fine-tuning (i.e., retraining of the deep learning model after transfer). We also tested the impact of picking deeper (and more expensive) models.

We published our main findings in our paper "*Knowledge transfer for melanoma screening with deep learning*" [Menegola et al., 2017a], from which we reproduced some parts to compose this thesis. The present author contributed to discussions, result evaluation, and writing of the published paper.

4.4.1 Experimental Proposal

In all experiments we adopted a deep learning framework for feature extraction of the skin lesion dataset, which fed a SVM classifier. We refer the reader to the published paper for further details regarding implementation.

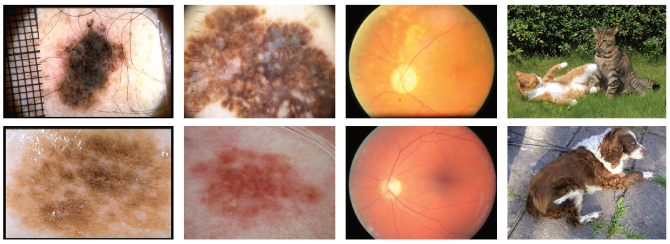
The datasets employed to train and test the target models (melanoma screening) were the Interactive Atlas of Dermoscopy (Atlas), and the ISBI Challenge 2016.

The sources datasets employed for the transfer (pre-training of the DNNs) were the Retinopathy dataset and the ImageNet Large Scale Visual Recognition Challenge 2012 dataset (ImageNet), containing 1M training images labeled into 1,000 categories [Deng et al., 2009]. Our use of ImageNet dataset was indirect — because training over it is so time consuming, we opted for employing pre-trained source DNNs.

Our design aimed to address two questions: 1) the impact of transfer learning, and

2) the impact of fine tuning (FT, that is, adjusting the network for the target model — melanoma identification — after transferring the source model).

For question 1), we evaluated the predictive model of the final classifier when using *no transfer* (Atlas from scratch), *transfer from Retinopathy* (Atlas from Retinopathy), *transfer from ImageNet* (Atlas from ImageNet), or *double transfer* (Atlas from Retinopathy from ImageNet). For question 2), we evaluated the predictive model of the final classifier when using *no FT* (only the final SVM decision layer is trained in the target models), or *with FT* (the target model is retrained for melanoma, before the last layer is stripped and the SVM decision layer is trained). The factors are shown in Figure 12.



(a) (b) (c) (d)

Figure 12 – Samples from datasets used in this experiment: (a) Atlas; (b) ISIC; (c) Retinopathy; (d) ImageNet. Each row shows a sample from a different class in the dataset. In this experiment, 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*. Image reproduced from Menegola et al. [2017a].

With a single exception, all protocols were based upon the VGG-M model proposed by Chatfield et al. [2014]. We also run a single comparison with the VGG-16 model [Simonyan and Zisserman, 2014] to evaluate the impact of that deeper (and more expensive) architecture.

In the experiments with transfer learning, we get the source networks pre-trained on ImageNet, train them from scratch on Retinopathy, or fine-tune on Retinopathy the model pre-trained on ImageNet. In the baseline (control) experiment without transfer learning, the networks are trained from scratch. In all networks, we ignore the output layer and employ 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 is involved (when we fine-tune or train networks from scratch) we employ the technique of *data augmentation*, which consists in creating randomly modified training samples from the existing ones. Data augmentation also brings the opportunity to rebalance the classes, alleviating another problem of medical tasks, where the positive class (melanomas) is usually much smaller than the negative (benign lesions). To accomplish both augmentation and balance, we only augment the minority classes (Melanoma, Malignant, and Basal Cell Carcinoma, depending on the experimental design).

	AUC	mAP	ACC	SE	SP
1st place	80.4	63.7	85.5	50.7	94.1
2nd place	80.2	61.9	83.1	57.3	87.2
3rd place	82.6	59.8	83.4	32.0	96.1
This work	80.7	54.9	79.2	47.6	88.1

Table 1 – Results on ISIC, with VGG-16, with transfer learning from ImageNet, with fine tuning. Baselines quoted from the ISIC 2016 competition website (competitors were originally ranked by mAP). The results are provided to show that the models evaluated here are realistic — in the sense that they are in the same ballpark of performance as current art. All numbers in %. Table reproduced from Menegola et al. [2017a].

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

- Malignant *vs.* Benign lesions: melanomas and basal cell carcinomas were considered positive cases and all other diagnoses were negative cases;
- Melanoma *vs.* Benign lesions: melanomas were positive cases while all other diagnoses were negative ones, removing basal cell carcinomas;
- Basal cell carcinoma *vs.* Melanoma *vs.* Benign lesions: here we have three classes, with all other diagnoses under a single Benign label.

For all designs we employed 5×2-fold cross-validation. Our splits were semi-random, making an effort to balance as much as possible diagnose distributions, to avoid unnecessary variability.

Our main 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. For the experiment with the ISIC dataset we also report the other measures employed in the competition. We show the results on ISIC for reference purposes, to demonstrate that the models being discussed here are in the same ballpark of performance as the current state of the art (Table 1).

4.4.2 Results and Analyses

The main results are in Table 2. Fine-tuning improves classification, both when transferring from the small-but-related dataset (Retinopathy), and when 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) leads to worse results than transferring from the general task of ImageNet, even in combination with the latter. That might indicate that transferring from very specific tasks poses special challenges for overcoming the specialization — even if the source and target tasks are somewhat related. The best protocol we found was to simply transfer from ImageNet, with fine-tuning. The comparison

Experimental Design	No Transfer	From Retinopathy		From ImageNet		Double Transfer
		no FT	with FT	no FT	with FT	with FT
Malignant *vs.* Benign	76.0	72.8	76.0	79.1	**82.5**	78.8
Melanoma *vs.* Benign	75.7	73.5	75.3	77.9	**80.9**	**80.9**
Melanoma *vs.* Carcinoma *vs.* Benign	73.0	71.4	72.8	79.4	**83.6**	81.8

Table 2 – Main results (AUC in %; FT: fine tuning). Surprisingly, transfer from another specific medical task (Retinopathy) is not effective, even if preceded by transfer from a general task (ImageNet). Fine-tuning has major impact and should be considered a necessity. The choice of labeling has a small and somewhat inconsistent impact, that might be due to chance. Table reproduced from Menegola et al. [2017a].

	AUC (%)		
Architecture	Mal×Ben	Mela×Ben	Mela×Carc×Ben
VGG-M	82.5	80.9	83.6
VGG-16	**83.8**	**83.5**	**84.5**

Table 3 – 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 reproduced from Menegola et al. [2017a].

	AUC (%)			
Experimental Design	Low	Medium	High	All
Malignant *vs.* Benign	93.7	82.5	58.8	82.5
Melanoma *vs.* Benign	93.0	79.6	56.6	80.9

Table 4 – 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 reproduced from Menegola et al. [2017a].

between DNN architectures shows that — as usually observed for image classification — a deeper DNN performs better (Table 3).

The experimental designs also showed differences in performance: in general it was easier 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 for those aggregate numbers might be simply that Basal cell carcinomas are easier to diagnose than Melanomas.

We show the results stratified by diagnose difficulty (as indicated by the Atlas itself) in Table 4. Those results show that low-difficult lesions can essentially be solved by current art with relatively high confidence, while for difficult lesions performance is still little better than chance.

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 (e.g., can another medical task show better results? Can another transfer scheme work?).

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).

The results stratified by diagnose difficulty suggest that current art can already deal with the lower and middle spectrum of difficulty, especially considering that human doctors' accuracies might be between 75-84% [Ali and Deserno, 2012]. On the other hand, difficult lesions appear *really* hard to diagnose.

4.5 Designing a powerful *melanoma classifier*

We participated in the 2017's edition of the "ISIC Challenge on Skin Lesion Analysis Towards Melanoma Detection". Based on previous results, we knew two facts: 1) deep learning is greedy for data, and 2) the best deep learning approach is a transfer learning scheme from a huge dataset. So, we based our approach in 3 main factors: data, models, and simple strategies that could enhance existing architectures.

We provide details about data collection, data usage, preliminary results, and intermediate analysis, in our report [Menegola et al., 2017b], from which we reproduce part of the text to compose this Section. The present author contributed to discussions, experimental design, experimental setup, result evaluation, and writing of the report of participation.

4.5.1 Experimental Proposal

We started to gather as much as data as possible, which lead us to use of datasets listed on Section 4.3, expect the Retinopathy, the ISIC Challenge 2018 and Other Sources. We also excluded images that could cause annotation clashes with the challenge (like images without a diagnosis from the ISIC Archive, and images marked as "atypical nevi" from PH2).

The Challenge provided official sets for training, validating and testing approaches. However, since we add external data to the training set (which was allowed), we also constructed and *internal validation set*, which better represented the training set and, hopefully, the unknown testing set.

We gathered the datasets during the competition, so we fully trained models using part of the datasets (ISIC Challenge, ISIC Archive, and Atlas) and other models trained

in all available datasets (with the exclusions explained above). We call those different amounts of data as *semi* and *deploy*, respectively.

The open question, however, was the choice of the deep learning architecture. We based our approach in two **models**: ResNet-101 [He et al., 2016] and Inception-v4 [Szegedy et al., 2017]. Both were state of the art and were available in multiple frameworks, pre-trained for ImageNet with good results. We fine-tuned each model in a 3-class scheme: melanoma *vs.* seborrheic keratosis *vs.* other lesions.

Since many other competitors may had similar approaches, we performed a series of strategies aiming to improve the models' accuracy. We established an initial agenda of hypotheses to validate. Omitting a few speculative ideas we did not have time to touch, those were:

1. Compare standard-resolution images (224 for VGG and ResNet) to double-resolution images;
2. Contrast different strategies of class- and sample- weighting during training;
3. Compare normal training schedule with some form of curriculum-learning;
4. Contrast different regimens of training and test augmentation;
5. Measure the impact of adding SVM as a final decision layer;
6. Attempt to use the patient data (age and sex) on classification;
7. Attempt different model optimizers;
8. Add different types of per-sample normalization;
9. Add a final meta-decision based upon multiple models (ensemble, stacking, etc.)

Our normal procedure would be to attempt an (incomplete) factorial design, at least for the factors where we expected cross-effects (e.g., depth × resolution × weighting × scheduling × augmentation). For the competition, however, there was no time for such level of rigor. We tested the hypotheses more or less sequentially, revisiting only those that seemed too surprising, and crossing only the effects for which we had a very strong expectation for interactions.

Failures: 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, and we performed no significance nor equivalence tests *at that time*. We sort the 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. We attempted several class- and sample- weighting schemes, both to compensate the unbalancing of the classes, and the reliability of the annotations. In one case we attributed weights inversely proportional to the frequency of the classes; in another case we combined those with weights that went from 1 to 3 ranging from "unknown follow-up"/"no follow-up" until "confirmed by histopathology" (we attributed 5 to the official dataset to give it extra weight). 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 in the results. We could not afford the *very* long fine-tunings (several weeks) recommended for Inception in some applications[6]. It is possible that in those super-long training scenarios early stopping with validation becomes important.

4. Patient data: we attempted different encodings for incorporating the patient data (age and sex) into the features, inserting them in the transition between the deep model and the SVM decision layer. The results were inconsistent, sometimes improving and sometimes worsening the results.

5. Curriculum learning: curriculum learning consists in careful scheduling of the training samples in order to present a "curriculum" of learning steps to the algorithm, instead of learning the samples at random (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 simply training with all images at once.

Success Factors: if most attempts disappointed, some definitely were valuable, as measured by both the internal and official validation AUCs. We have not, for the moment, a factorial analysis to quantify the contributions, but we sort the list placing first the factors we believe helped the most:

[6] https://github.com/tensorflow/models/tree/3be9ece9574d7bac07704e43705741d9af1de1e6/im2txt#fine-tune-the-inception-v3-model

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

2. Data augmentation: from experience, we knew that training with data augmentation is critical (i.e., applying random transformations: croppings, flippings, etc. on the images before using them in the network) and made all attempts with it. 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 a 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 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. Surprisingly, going one step further and dividing the pixels by the standard deviation gave worse results than no per-image normalization at all.

4. Stacking models and meta-learning: fusing the decision of several models gave, almost always, better results than just using the single best model, even when using simple schemes, like averaging the probabilities among the models for a given sample. However, 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.

4.5.2 Results and Analyses

Figure 13 shows a subset of 48 out of more than a hundred models we evaluated (most experiments were too quick-and-dirty to allow inclusion in the plot). 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 who 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

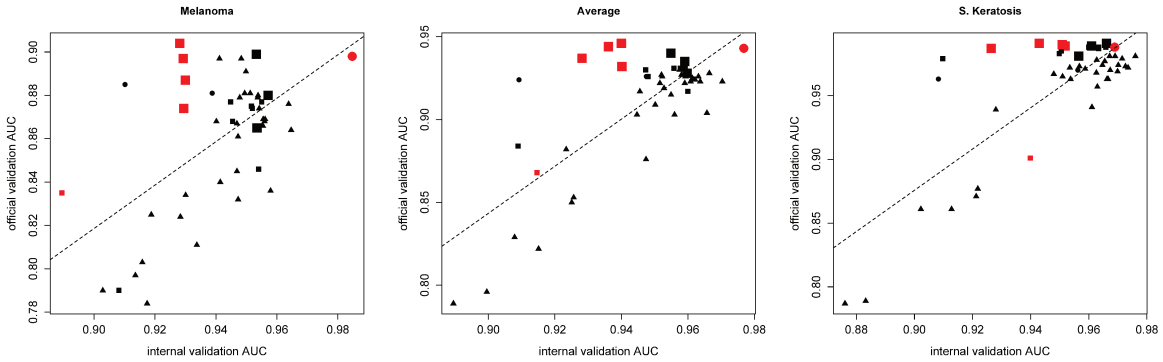


Figure 13 – A visual panorama of our experiments. The circles are ResNet-101 Models, the triangles and squares are Inception-101 models (without and with per-image normalization respectively). Black and red indicates training in the deploy and semi datasets respectively. Large symbols indicate the models chosen to compose the stacking in the final submission. The dashed line is the regression between the internal and the official AUCs. The models are the same in the three graphs, but the metrics change from Melanoma, Average, and Seborrheic Keratosis AUC from left tor right. Those were only a subset of the experiments, 48 out of more than a hundred models we attempted. Image reproduced from Menegola et al. [2017b].

tasks, but in the end we decided to pick the same set of models for both tasks and hope the meta-learning layer would do the adjustments.

The meta-learning consisted in, for each sample, concatenating the decisions of each chosen model and using this as feature 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 AUC.

We attempted several small variations for the meta-learning; the most successful employed an aggressive augmentation scheme: each training sample (from our internal validation set) was evaluated thrice by each model, allowing to create a large number of combined replicas for training. The same procedure was applied for testing (on the samples from the official validation and official test sets). In both cases, we employed average pooling to combine the replicas.

Our meta-learning approach consisted in, for each sample, concatenating the decisions of 7 models and using them as feature vectors for two binary SVMs (melanoma-vs-all, keratosis-vs-all). The seven base models are six based on Inception and one based on ResNet.

The approach described above took us to the first place on the "melanoma *vs.* all" subtask, which is the most relevant one, regarding screening purposes.

Symbol	Factor	Levels
a	Model	ResNet-101 v2 *versus* Inception v4
b	Train dataset	Train split of ISIC Challenge 2017 *versus* Full: Level 1 + ISIC Archive + U. of Porto PH^2 + U. of Edinburgh Dermofit
c	Input resolution	Pre-augmentation resolution — 299×299 pixels (305×305 if using segmentation) *versus* 598×598 pixels
d	Train augmentation	TensorFlow/Slim's default *versus* Level 1 + rotations = on, fast_mode = off, minimum_area = 0.20
e	Input normalization	TensorFlow/Slim's default *versus* Subtract mean of samples' pixels
f	Segmentation	No segmentation information *versus* Segmentation pre-encoded at input
g	Training length	Short (about half the length of Full) *versus* Full (30k batches for ResNet / 40k batches for Inception; batch_size = 32)
h	SVM decision layer	Absent *versus* Present
i	Test augmentation post-deep	No (decision on single non-augmented sample) *versus* Yes (decision on average of 50 random-augmented samples)
j	Test dataset	Split of Train/Full *vs.* Validation of ISIC Chall. '17 *vs.* Test of ISIC Chall. '17 *vs.* EDRA/Dermoscopic *vs.* EDRA/Clinical
t	Transfer learning	Training from scratch (weights initialized at random) *vs.* Transfer from ImageNet (checkpoint published by Tensorflow/Slim)

Table 5 – Factors in our experimental designs, with corresponding levels. Table reproduced from Valle et al. [2017]

4.6 How to Extract *Greater Performance* From Deep Models?

Our participation in the 2017 ISBI Challenge raised some questions of what are the main aspects that contribute to improving classification rates of methods based on deep learning. That motivated the proposal of an experimental design to eliminate any doubt. We explore ten choices often faced to design deep networks: model architecture, training dataset, image resolution, type of data augmentation, input normalization, use of segmentation, duration of training, additional use of SVM, test data augmentation and testing set (Table 5).

Most of those factors are not particular to melanoma detection, but are relevant for all image classifiers using deep learning. However, a preoccupation with resolution (c), augmentation customization (d), and segmentation (f) makes more sense for melanoma detection — or at least for medical images in general — than for general-purpose tasks, like ImageNet.

We perform a two-level full factorial experiment, for five different test datasets, resulting in 2560 exhaustive trials, which we analyze using a multi-way ANOVA.

Here we summarize the main findings and discuss the new perspectives brought by our results. The details are present in one of our publication [Valle et al., 2017], which parts of the text are reproduced in this Section. The present author contributed to discussions, experimental designing, result evaluation, and writing of the paper under review.

Type	Melanoma	Nevus	Keratosis
ISIC Challenge 2017 train split	374	1372	254
Full train (composition of datasets)	1227	10124	710
Internal test split from full	135	3129	89
ISIC Challenge 2017 validation split	30	78	42
ISIC Challenge 2017 testing split	117	393	90
EDRA Atlas of Dermoscopy (each version)	518	1154	95

Table 6 – Summary of the train and test sets. Table reproduced from Valle et al. [2017].

4.6.1 Experimental Proposal

The main experimental design was **a two-level full factorial design** for nine of the ten factors mentioned above (a–i), for each one of the five test datasets (j), resulting in $2^9 \times 5 = 2560$ treatments evaluated. We run a second experiment to evaluate **the impact of transfer learning**, evaluating seven factors (a–e, g, i, t), and fixing (f) as no segmentation and (h) as SVM layer absent, resulting in $2^8 \times 5 = 1280$ treatments evaluated. 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.

The analysis for both experiments was a classical multi-way ANOVA, in which the test datasets entered as one of the factors. That choice highlights our aim to make decisions that generalize across datasets, in contrast to maximizing the performance for a particular dataset.

We used all skin lesion datasets listed in Section 4.3, except the IRMA Dataset and the ISIC Challenges 2016 and 2018.

Data sources affect the train and test datasets. For the train dataset (factor b), we contrasted (1) using only the official train split of the ISIC Challenge 2017 dataset, to (2) joining the train split of the ISIC Challenge, the ISIC Archive, the Dermofit Library, and the PH2 Dataset and extracting from that full dataset a train split. For the test dataset (factor j), we contrasted (1) an internal test split extracted from our full dataset; (2) the official validation split and (3) the official test split of ISIC Challenge 2017; (4) the dermoscopic images and (5) the clinical images of the EDRA Interactive Atlas of Dermoscopy. Table 6 summarizes the final assembled sets.

Segmentation was used only as an ancillary input for classification (factor f). For the ISIC Challenge 2017, we had used a segmentation network based on the work of Ronneberger et al. [2015] and Codella et al. [2015]. For this work, we streamlined that model, reducing the number of parameters, removing the fully-connected and Gaussian-noise layers, and adding batch-normalization and dropout layers. The new model[7] is faster to train and occupies much less disk space. We trained the segmentation models on the

[7] https://github.com/learningtitans/isbi2017-part1

same images as their corresponding classification models.

Skin lesion segmentation is a very challenging task by itself, posing some challenges like handling with inter-annotator agreement [Ribeiro et al., 2019]. Because of the lack of literature consensus on how to use segmentation for melanoma detection, we opted for schemes with minimal changes to both data and networks. We pre-evaluated two candidates: pixel-wise multiplying the input RGB images by the segmentation masks *versus* pre-encoding the four planes (R, G, B, and mask) into three planes, keeping the rest of the networks unchanged. For the full design, we only considered the latter, which appeared more promising on those preliminary tests.

Pre-encoding the masks required slightly adapting ResNet and Inception, by adding the pre-encoding adapter layers. For both ResNet/Inception we added three convolutional layers before the input, two layers with 32 filters, and a third with 3 filters. All convolutional layers used 3×3 kernels and stride of 1. Since ResNet-101-v2 and Inception-v4 models require input images of 299×299 pixels, the adapter layer took 305×305-pixel images, to account for the 2 border pixels lost at each convolutional layer.

Most of the time, our full factorial designs are too costly to use — thus our next set of experiments, **exploring ensemble techniques**, helps in more practical situations. We evaluated a straightforward ensemble, which just pools the decision of several classifiers, and showed that it provides very good performances, without the costs of a full design.

We also simulated the most common procedure employed by researchers and practitioners: **sequential optimization of hyperparameters**, in which one starts from a given configuration of hyperparameters, selects one of them to evaluate, commits to the best results, and proceeds to evaluate the next. Although such procedure is very fast (it allows optimizing the nine factors our main design in just 18 experiments), it is sub-optimal in comparison to ensembles.

Finally, we showed that the customary procedure of optimizing the **hyperparameters in the same test set** used to evaluate the technique leads to overoptimistic results in both the ensemble and the sequential design.

4.6.2 Results and Analyses

Here we describe the results of each aspect described in previous subsection (in **bold**). The reader interested in particular results can go direct to the highlighted subparts of the text (with ***bold and italic***).

Main Results: as explained, the main experiment was a full factorial design with nine two-level factors (a–i), and five test datasets (factor j). We used a classical multi-way ANOVA with the mean AUC for melanoma and keratosis as the measured outcome (with the small technicality of taking the logit of that measure, since, when working with rates,

			Explanation (%)		Best AUC (%)		Worst AUC (%)	
	Factor	p-value	Absolute	Relative	Treatment	Mean	Treatment	Mean
a	Model architecture	<0.001	0	1	resnet	84	inception	83
b	Train dataset	<0.001	5	46	full	85	challenge	81
c	Input resolution	<0.001	1	5	598	84	299–305	82
d	Data augmentation	0.17	0	0	default	83	custom	83
e	Input normalization	0.001	0	0	default	83	erase mean	83
f	Use of segmentation	<0.001	0	2	no	84	yes	83
g	Duration of training	0.003	0	0	full	83	half	83
h	SVM layer	<0.001	0	4	no	84	yes	83
i	Augmentation on test	<0.001	1	12	yes	84	no	82
j	Test dataset	<0.001	75		full.split	96	edra.clinical	66
a:b		<0.001	1	8	inception/full	86	inception/challenge	80
a:f		<0.001	0	2	resnet/no	84	inception/yes	82
b:e		<0.001	0	2	full/default	86	challenge/default	80
b:j		<0.001	2		full/full.split	98	chall/edra.clinical	63
h:j		<0.001	0		no/full.split	97	yes/edra.clinical	65
i:j		<0.001	0		yes/full.split	97	no/edra.clinical	65
a:b:d		<0.001	0	2	inception/full/custom	86	inception/challenge/custom	78
a:d:e		<0.001	0	2	resnet/custom/default	85	inception/custom/default	81
a:f:j		<0.001	0		resnet/yes/full.split	97	inception/yes/edra.clinical	65
b:d:e		<0.001	0	1	full/custom/default	86	challenge/custom/default	79
c:e:f		<0.001	0	1	598/default/no	86	299–305/default/yes	82
	Residuals	—	12					

Table 7 – Selected lines from the 176-line ANOVA table; most of the omitted lines (126) had p-values ≥ 0.05. Absolute explanation based on η^2-measure, relative explanation ignores residuals and choice of test dataset (j). Table reproduced from Valle et al. [2017].

the logit helps to fulfill ANOVA's assumption of Gaussian residuals). We considered all main effects, and up to 3-way interactions. We considered higher-order interactions unlikely and assigned them to the residuals.

Table 7 shows a summary of main experiment's ANOVA, with the symbols for the factors and interactions on the first column, and the names of the main factors on the second. The remaining columns show the outcomes of the test. The most important columns are *p-value*, which measures statistical significance, and *explanation (%)*, which measures effect-size/explanatory power.

We present the absolute explanation (considering the entire table) for reference, but our analysis is focused on the relative explanation, which ignores the choice of the test set (j) and the residuals. The reason for ignoring those is that they are not actual choices for designing a new model; therefore, relative explanations indicate better the relative importance of choices to practitioners.

The original full table contained all main effects, and up to 3-way interactions. However, not surprisingly, 126 of the resulting 176 lines were non-significant interactions, which were omitted here. We also left out those interactions with relative explanations lower than 1%, even if significant. With the notable exception of the customized data augmentation (d), all main effects were significant, but most of their relative explanations were small.

The analysis of the relative explanation shows an unsurprising, but still disappointing result: the performance gains (b) are almost wholly due to the usage of more data. Other than data, the most important factor was the use of data augmentation on

test (d). We performed it, as usual, by taking the test image, generating a number (in our case, 50) of augmented samples exactly like in training, collecting the prediction for each of the samples, and pooling the decisions (in our case, by taking the average prediction). Although not surprising for the literature of deep learning, that finding is relevant for the literature of melanoma detection, where many works still forgo augmentation in the test.

Most of the findings tended to confirm the (limited) observations we made during the ISIC Challenge 2017, with two notable exceptions. Input resolution (c), which we deemed unimportant during the challenge, turned out to have a non-negligible effect. That result is particularly interesting, because we used a very rough form of augmented resolution, by inputting high-resolution images to the augmentation engine, but still feeding normal-resolution crops to the network. On the other hand, the use of an SVM decision layer (h), which we considered advantageous during the Challenge turned out to have a large-effect... only negative! Globally, ANOVA shows it is better *not* to use the SVM.

Normalization (e) and training duration (g) showed tiny (<1%), but still significant positive effects. The choice for those factors must consider their very different costs: adding normalization costs next to nothing, both in implementation complexity and in training time. Training duration doubled the already many-hours-long training times.

As usual, most of the interactions were not significant, and even the ones that were, had effect sizes too small to be worth noting. A notable exception was the interaction between model architecture and train dataset (a:b), whose 8% of relative explanation was bigger than most main effects. Model choice alone favors the simplest ResNet over Inception, but the combination of Inception with the full dataset is so advantageous that it offsets that effect. We had already observed, informally, this synergy between more data and deeper models during the Challenge.

The most disappointing result was the use of segmentation, which was more than unhelpful, harmful. This result, however, is contingent on our choice for adding segmentation to classification.

Correlations of Metrics and Testing Sets: we performed an additional correlation analysis with the full factorial experiment (Figure 14), to highlight the correlations (a) among results on different test datasets; and (b) among different metrics. To keep the scatter plots directly interpretable, instead of taking the logit of the rates, we dealt with the non-linearity by using Spearman's ρ instead of Pearson's r as correlation measure.

The correlogram on Figure 14(a) considers, as the ANOVA, the mean melanoma/keratosis AUC. The test dataset names appear in the diagonal, along with the maximum and minimum AUCs obtained for the 512 variations of the full design on that dataset. The scatter plots in the upper-triangular matrix follow the usual construction for correlograms. The lower-triangular matrix displays the Spearman's ρ's: the mean

estimate appears as the printed numeral and as the area of the solid circle; the bounds of the 95%-confidence interval appear as the area of the internal and external dashed circles. Negative correlations appear in red.

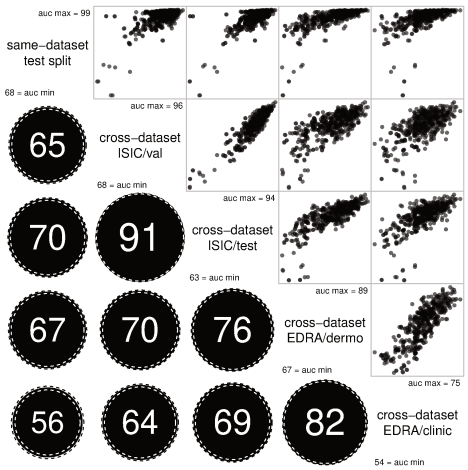
The correlation between different test datasets is far from perfect. That is, perhaps, obvious, but must be stressed, since it reveals that *naively* hyperoptimizing a model on one test set will not necessarily generalize to other data. The relationship between splits of different datasets is more subtle. Note how the correlation between the validation and the test splits of ISIC 2017 Challenge, and the dermoscopic and clinical splits of EDRA have the highest correlations. This suggests that results measured on splits of the same dataset may not wholly generalize over data of the same type obtained on different conditions. Both phenomena show how hyperoptimizing on test gives unwarranted advantages, leading to overoptimistic assessments.

The correlogram on Figure 14(b) considers only the results for the test split of the ISIC Challenge. Different metrics appear in the diagonal: average precision, area under the ROC curve, sensitivity (true positive rate), and specificity (true negative rate), for both melanoma and keratosis. The interpretation of the plots, numerals, circles, and colors is the same as above.

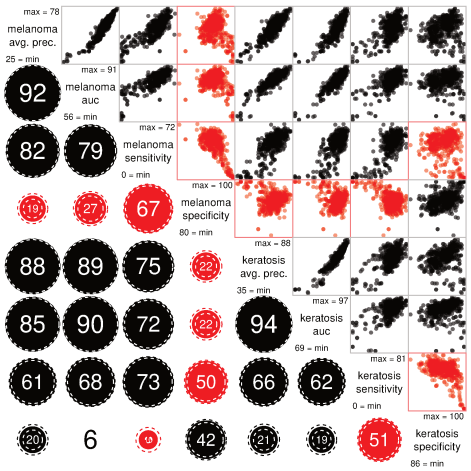
This correlogram is interesting for showing that many metrics have correlations that are not that big. Particularly noteworthy is the specificity, which has not only a negative correlation with sensitivity (as expected), but also a negative or very small correlation with *most of* the other metrics.

The Impact of Transfer Learning: we run a second full factorial design, with seven of the ten factors of the main experiment (a–e, g, i, j), fixing factors (f) and (h), and adding a factor to evaluate the presence *versus* absence of transfer learning (factor t). The new factorial design, with $2^8 \times 5 = 1280$ treatments, shows transfer learning as critical for performance: it explains (favorably) 14.7% of the absolute variation, and a whopping 62.8% of the relative variation of performance (computing those metrics the same way as the in the main experiment, i.e., excluding the residuals, and the choice of test dataset and its interactions from the relative variation), with high significance (p-value below 0.001). We omit the ANOVA table for concision. Those results reinforce previous findings on the importance of transfer learning for melanoma detection [Menegola et al., 2017a].

The Sequential Procedure to Design a Model: the sequential procedure tries to avoid the costs of full factorial designs, whice are way too expensive for the majority of situations. The sequential procedure relies on taking a single factor to optimize, and performing a couple of experiments on that factor alone, keeping all others fixed (starting from a combination considered reasonable). Once a factor is decided, one commits to it and takes the next to optimize, until the procedure is complete. Here we evaluated the impact of such approach.

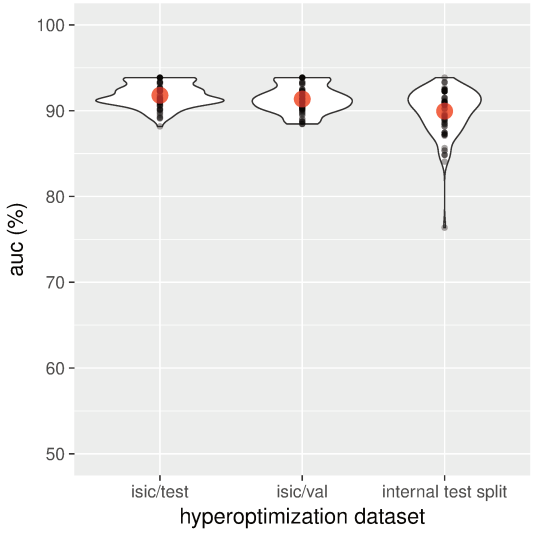


(a) Correlogram of mean melanoma/keratosis AUC across test datasets.

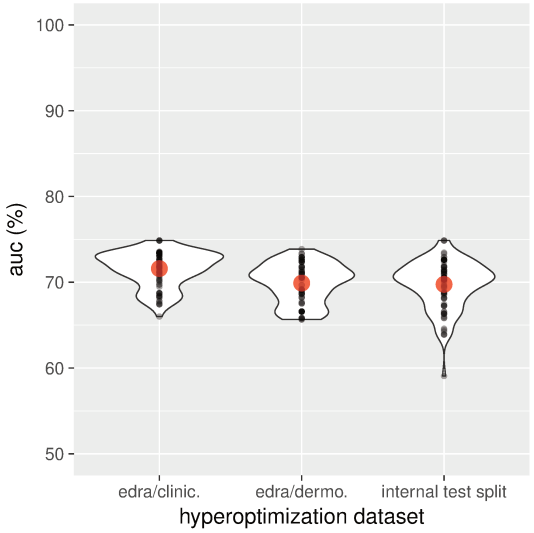


(b) Correlogram of metrics on ISIC 2017 Test dataset across metrics.

Figure 14 – Correlograms with pair-wise correlation analyses. Sets appear on the diagonal; upper matrices show the scatter plots, and lower matrices show the Spearman correlation of each pair of sets. On lower matrices, numbers and solid circles' areas show the mean estimates, and dashed circles' areas show the 95%-confidence bounds. Non-significant estimates appear without the circles. All numbers in %, negative correlations are in red. Image reproduced from Valle et al. [2017].



(a) Simulated sequential design on ISIC test split.



(b) Simulated sequential design on EDRA Atlas clinical images.

Figure 15 – Simulation of the sequential optimization of hyperparameters, considering all nine factors (a–i) as the main full factorial. Factors optimized on the dataset shown on the horizontal axis, and performance (mean melanoma/keratosis AUC) measured on the dataset indicated on the captions. For each case, we run 100 simulations, with random optimization sequences and starting points. The violin plots show the kernel density estimation of the actual data (black dots) and the large red dot shows their mean. Image reproduced from Valle et al. [2017].

We take, at random, both the starting treatment and the sequence of factors to test. For factors not yet optimized, the level is given by the starting treatment. Each factor is optimized in turn, by comparing the performance of the alternative treatments on the full-factorial data of a chosen hyperoptimization dataset. The outcome of a single simulation is the performance of the optimized treatment on a chosen measurement dataset. We use the mean keratosis/melanoma AUC as the performance metric.

Figure 15 shows the results for pairs of hyperoptimization× measurement datasets, where we perform 100 simulations for each pair. The actual measurements appear as black dots, and the violin plots show their estimated density, while the big red dot shows their mean. The most notable observation is the (unrealistic) advantage of hyperoptimizing and measuring on the same dataset: not only do we get higher averages, but also a smaller variability. The advantage of hyperoptimizing and measuring on splits of the same dataset is more subtle, but present.

The expense of the full factorial design, the instability of the sequential procedure, and the limited correlation of performances across datasets seem to leave few options to practitioners. Fortunately, single-model schemes are seldom used today, and ensembles of several models help to alleviate those issues.

Ensemble Approaches: we simulated different ensemble strategies, by pooling the predictions of models present in our full design. We evaluate three pooling strategies: average, max, and extremal. Average- and max-pooling work as usual. Extremal pooling takes, from the list of values being pooled, the value most distant from 0.5 — it may be seen as an "hypothesis-invariant" max-pooling. In all cases, after pooling, we re-normalize the probability vector to ensure it sums up to one. Half the models in the full design entered as candidates, and we discarded in this experiment the models with the SVM layer, due to issues in making their probabilities commensurable with the deep-only models.

Figure 16 shows the main results. Average-pooling was, by far, the best choice for pooling the decision. Such clear-cut advantage came as a surprise for us, as max-pooling often outperforms average-pooling in related tasks. If no other information is available, simply average-pooling randomly selected models is a reasonable strategy.

The use of dozens — even hundreds — of models may sometimes be justified in critical tasks (like medical decisions), but training and evaluating so many deep networks is cumbersome. Fortunately, as Figure 17 shows, a handful of models seem to work just as well. The results shown here are the "good news" part of this paper: we can escape the expense of the full factorial design, and the instability of the sequential designs, by averaging a dozen or so models with parameters chosen entirely at random — although the random ensembles start very unstable, they soon converge to a reasonable model, in average and variability. If we decide to perform a full factorial, there is good news too: the best models learned in one dataset seem to be informative to compose the ensembles

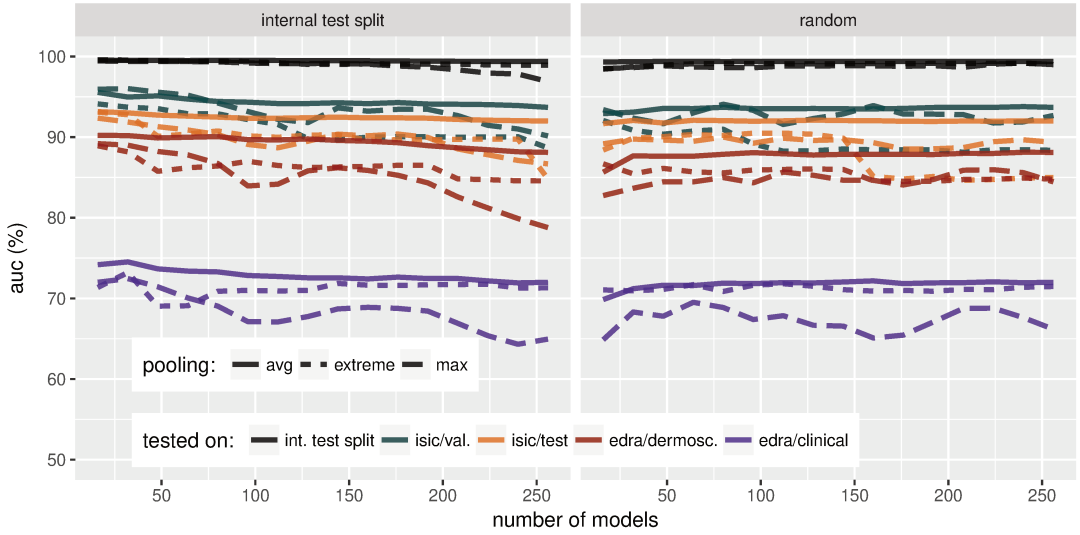
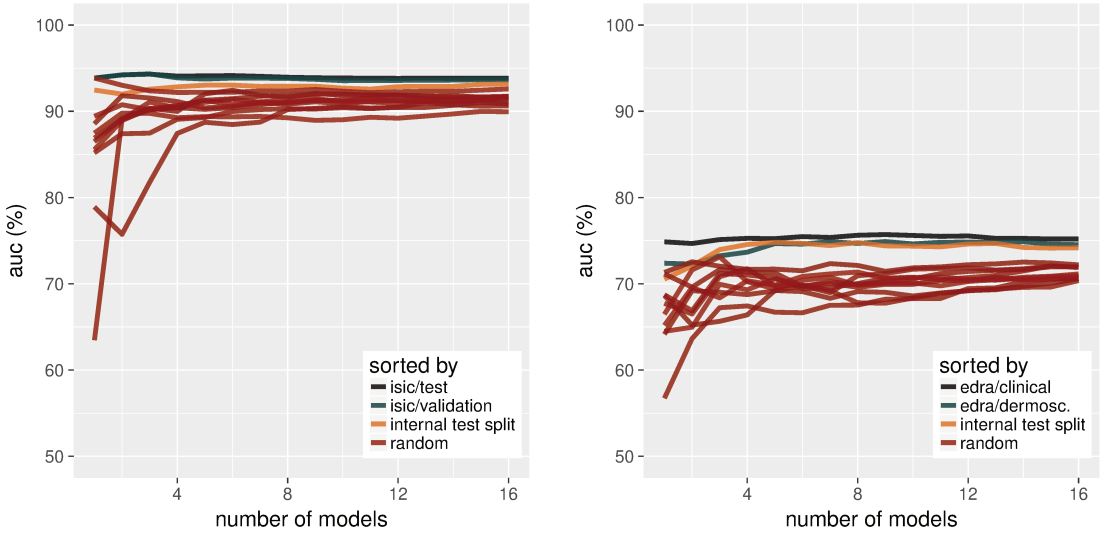


Figure 16 – Evaluation of ensemble strategies, by pooling the prediction of a given number of partial models using average-, max-, or extremal-pooling. Left: cumulative effect of adding partial models, starting with the best (as evaluated by the internal test split). Right: same plot, with models randomly shuffled. (Best viewed in color.) Image reproduced from Valle et al. [2017].

in other datasets, allowing to get top performances with very small ensembles.

Here, again, the unfair advantage of optimizing (selecting the models for the ensemble) and measuring performance on the same dataset appears. The advantage is small but systematic for the test split of ISIC (Figure 17(a)); it is much more apparent for the challenging collection of clinical images of EDRA Atlas (Figure 17(b)).



(a) Ensembles on ISIC test split.

(b) Ensembles on EDRA Atlas clinical images.

Figure 17 – Detailed analysis of ensembles, contrasting the dataset used to choose the models (shown as different curves) in order to optimize the results for a measurement dataset (indicated in captions). We sampled 10 different random ensembles. (Best viewed in color.) Image reproduced from Valle et al. [2017].

4.7 New Perspectives For *Skin Lesion Analysis*

Motivated by our good performance in the ISIC competition of 2017 as well as the discoveries described in Section 4.6, we decided to participate in the 2018 edition of the ISIC competition.

We participated in the three tasks of the Challenge, but here we describe the experiments related to image classification. The other tasks and detailed explanations can be found at our report [Bissoto et al., 2018a], from which we reproduce part of the text to compose this Section. The present author contributed to discussions, activities organization, and writing of the report.

The experiments conducted in the previous months allowed the creation of a robust framework, aimed at increasing the performance of computational methods for automatic melanoma screening. However, the main task of the competition is no longer the identification of melanomas amongst other lesions, and it has become a multiclass task, which requires discrimination of multiple types of skin lesions.

It is, therefore, an even more complex and even more demanding problem, particularly for the minority classes. To improve our chances, we also introduced two original contributions — synthetic lesions generation and stronger data augmentation approaches — to boost the models training. Such contributions will be detailed next.

4.7.1 Experimental Proposal

In previous work, we showed that the training set size responds by almost 50% of the variation on the prediction power of the classifier [Valle et al., 2017]. The freedom to use external sources enabled us to gather more data to boost our models. Therefore, in addition to the official basis of the competition 2018, we also use the data ISIC Archive, Atlas, Dermofit, PH2 and Other Sources (Section 4.3), ending with 30,324 images (with diagnosis label).

After picking a dataset, we divided it the into 3 splits, for each task: 10% for holdout (for our internal model selection) and the remaining 90% for training. The training split was further divided into five 10%-validation/90%-training different splits (at random, not using cross-validation folds). We considered case numbers, aliases, and near-duplicates in the split division, to minimize contamination across splits.

We used the holdout sets to select the models. We used the metrics observed in the holdout sets to identify strong release candidate models and/or good bets for a meta-learning phase. Although the official validation data was very limited on this year's Challenge, we still used its scores as ancillary estimates. The exact datasets and splits,

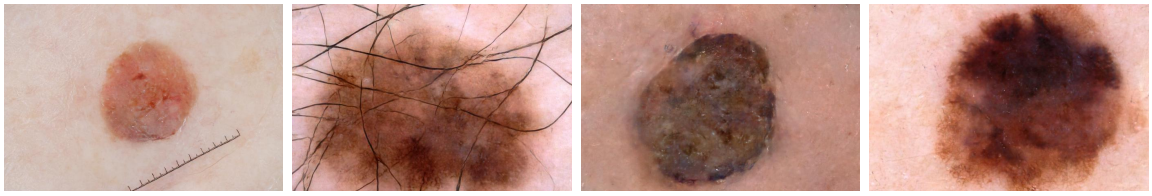


Figure 18 – Samples of high-definition, visually-appealing, clinically-meaningful synthetic skin lesion images. All samples are synthetic. Image reproduced from Bissoto et al. [2018b].

for each task, are listed, image by image, in our **code repository**[8].

For this year we took advantage of our recent results regarding new approaches for data augmentation: (a) image processing of real skin lesion images [Perez et al., 2018], and (b) synthetic skin lesions using Generative Adversarial Networks (GANs) [Bissoto et al., 2018b].

In work (a), we investigated the impact of 13 image processing-based scenarios of data augmentation for melanoma classification. Scenarios include traditional color and geometric transforms, and more unusual augmentations such as elastic transforms, random erasing and a novel augmentation that mix two different lesions. Using our participation on ISIC Challenge 2017 (with Inception-v4) with as baseline, we observed similar performance using the new data augmentation methods, but without using external data. That is, the image processing data augmentation methods were equal to the performance of the model trained with external data (which we know that has a huge impact on the classifier prediction power). Among all experiments and scenarios, scenario J (random crops, simulation of camera distortions, random horizontally and/or vertically flips and saturation, contrast, brightness and hue modifications by random factors — please refer to Perez et al. [2018] for details) leads to better performance and was the one introduced in the experiments of the competition (only in Task 3).

In work (b), we created fake high-resolution (1024×512 pixels) skin lesion samples, aiming to extend the training set artificially. To do that, we used GANs to teach the network the malignancy markers and also incorporating the specificities of a lesion border. Please refer to Bissoto et al. [2018b] for details. Figure 18 shows some examples of the generated images.

We used the synthetic images only on Task 3 (on the two submissions using external data). We added the synthetic images to the training/training splits (never to the holdout or to the training/validation splits) keeping a 1:1 per class proportion (i.e., one synthetic image for each real image in each lesion class).

We trained three different CNN architectures: Inception-v4 [Szegedy et al., 2017], ResNet-152 [He et al., 2016], and DenseNet-161 [Huang et al., 2017], all pretrained on ImageNet dataset. We fine-tuned the networks on three datasets: full, only, and full+synthetic

[8] https://github.com/learningtitans/isic2018-{part1,part2,part3}

augmentation.

To deal with dataset imbalance, we set the optimization goal to a class-weighted cross-entropy, with the weights calculated by dividing the frequency of the most common class by the frequency of each class.

We performed online data augmentation as described in Perez et al. [2018] (scenario J). We applied the transformations to the validation (single replica), holdout (32 replicas), and final test (128 replicas), taking the decision as the average of the replicas.

4.7.2 Results and Analyses

Our three submissions were (1) XGBoost ensemble of 43 deep learning models; (2) average of 8 best deep learning models (on the holdout set) augmented with synthetic images[9] and (3) average of 15 deep learning models trained only with Challenge data. Our final results on the official testing set were, respectively, 0.732, 0.725 and 0.803 for the normalized multi-class accuracy. Also, our positions of each submission were, respectively, 32th, 39th and 9th among 141 submissions.

We are very excited to see the ISIC Challenge as a continuing event, since we consider such initiative as pivotal for the development of our research area. Until recently, making comparisons across different approaches for skin lesion analysis was essentially impossible, due to difficulties of code and data sharing, and lack of standardized evaluation metrics and datasets [Fornaciali et al., 2016]. We also acknowledge the importance of keeping the testing set secret until all evaluations were over, preventing, thus, subtle methodological errors that inflate the performance evaluation of models [Valle et al., 2017; Salzberg, 1997].

Despite the diversity of skin lesion types and their dermatological importance, we asked ourselves whether making the classification task (Task 3) so fine-grained was really necessary, especially given the huge class imbalance. We fear that confusion among very small classes (e.g., Benign Keratosis and Actinic Keratosis) will bring much noise to the evaluation. In our current research, we are still focusing on coarse-grained melanoma/non-melanoma screening/triage classifiers — and we notice that real-world performances even for such coarse-grained procedures are still far from ideal.

4.8 Deep Learning Requires Data: Is Our Data *Flawless*?

Knowing the value of the data for machine learning, we decided to investigate the potential of the already existing datasets for the development of robust solutions. Our

[9] N.B. that approach is wrongly named as an average of *15 models* on the official leaderboard

objective was to analyze how the computational models take advantage of the information provided by the images, as well as to assess the quality of such bases.

Due to the scarcity of good-quality, annotated skin lesion images, two datasets dominate research on automated skin lesion analysis: the Interactive Atlas of Dermoscopy and the ISIC Archive (Section 4.3). The problem of having so few, relatively small datasets dominating much of research in automated skin analysis, is the risk of datasets biases. If bias is present even in bigger and more diverse datasets [Torralba and Efros, 2011] like ImageNet [Russakovsky et al., 2015], it is naive to think it is not present in the smaller and harder to obtain skin cancer datasets, where we lack works identifying the possible sources of dataset bias.

We also know that there are visible artifacts introduced during the image acquisition process (e.g., dark corners, marker ink, gel bubbles, color charts, ruler marks, skin hair) [Mishra and Celebi, 2016] that could inflate models performances due to spurious correlations. Despite being impossible to eliminate wholly, it is crucial to understand bias and its sources to improve our image acquisition processes and deep learning models further.

Our hypothesis is: if we hide the lesion information from the networks, can it still learn patterns that help differentiate benign from malignant lesions? We believe that when a model learns to classify malignant lesions by analyzing only the skin —without information on the borders, biological markers or lesions' diameter— it strongly relies on patterns introduced during image acquisition and general dataset bias.

We published our main findings in our last paper [Bissoto et al., 2019], from which we reproduced some parts to compose this thesis. The present author contributed to discussions, part of the experiments, and writing of the published paper.

4.8.1 Experimental Proposal

For our experiments, we select only the dermoscopic samples, remove "duplicates" (some medical cases have multiple images), and include only the classes present in the dataset of task 2 of 2018 ISIC Challenge (melanoma, nevus, and seborrheic keratosis). Those alterations result in a dataset containing 872 images.

We firstly employ the 7-point checklist [Argenziano et al., 1998], a score-based medical algorithm, to verify bias in the Atlas dataset. This way we can isolate the neural network's learning capabilities. Dermatologists use attribution pattern analysis to diagnose malignant cases. The 7-point medical algorithm assigns a score to each of the dermoscopic attributes (see Table 8). The medical practitioner needs to accumulate the scores over the detected present attributes. If this score surpasses a threshold, the lesion is assigned as a melanoma. We use the 7-points checklist score instead of other medical

#	Criterion	Definition	7-point score
	Major criteria		
1	Atypical pigment network	Black, brown, or gray network with irregular meshes and thick lines	2
2	Blue-whitish veil	Irregular, confluent, gray-blue to whitish-blue diffuse pigmentation	2
3	Atypical vascular pattern	Linear-irregular or dotted vessels not clearly combined with regression structures	2
	Minor criteria		
4	Irregular streaks	Irregular, more or less confluent, linear structures not clearly combined with pigment network lines	1
5	Irregular pigmentation	Black, brown, and/or gray pigmented areas with irregular shape and/or distribution	1
6	Irregular dots/globules	Black, brown, and/or gray round to oval, variously sized structures irregularly distributed within the lesion	1
7	Regression structures	White areas (white scarlike areas) and blue areas (gray-blue areas, peppering, multiple blue-gray dots) may be associated, thus featuring so-called blue-whitish areas virtually indistinguishable from blue-whitish veil	1

Table 8 – The 7-Point Checklist [Argenziano et al., 1998]. Seven features compose the checklist, divided into major and minor criteria. The major criteria are associated with 2 points each one; the minor criteria are associated with 1 point each one. The simple addition of the individual scores a minimum total score of 3 indicates a melanoma, whereas a total score of less than 3 suggests a non-melanoma.

algorithms because it was available as metadata of the Atlas dataset. It achieves 91.7% AUC over all selected Atlas samples (see Figure 19).

To help us to understand that result, we adopted the melanoma classification benchmark [Brinker et al., 2019] to measure the expected performance for dermatologists, in an unbiased scenario. This benchmark is the result of a study with 157 German dermatologists to be a reliable benchmark for artificial intelligence algorithms. Brinker 's procedure were to send an electronic questionnaire to dermatologists containing 100 dermoscopic images (80 nevi and 20 biopsy-verified melanoma) randomly chosen from the ISIC Archive, asking for their evaluation. The AUC achieved by dermatologists for dermoscopic images (which is the case for our Atlas set) is 67%.

The huge gap between the 7-point checklist performance with the melanoma classification benchmark reveals it is biased due to the characteristics and educational objectives of the Atlas dataset. Low and medium difficulty cases selected to compose the dataset are probably hand-picked to be good examples to teach new medical practitioners to identify and classify dermoscopic attributes, while hard cases are exceptions to the pattern-based analysis.

Next, we try to find the source of bias, by gradually destructing clinical-meaningful information from the images, and assessing the network's performance on them.

To accomplish our goals, we propose destructive actions in the target datasets (Figure 20). First, we introduce our ideas to exploit the deep neural network learning capabilities.

We use the same network architecture and hyperparameters for all experiments. We employ an Inception-v4 network [Szegedy et al., 2017], widely used for computer vision, and well-established for skin lesion analysis. We fine-tune the ImageNet [Russakovsky et al., 2015] pre-trained network to the target dataset. We resize the input images to

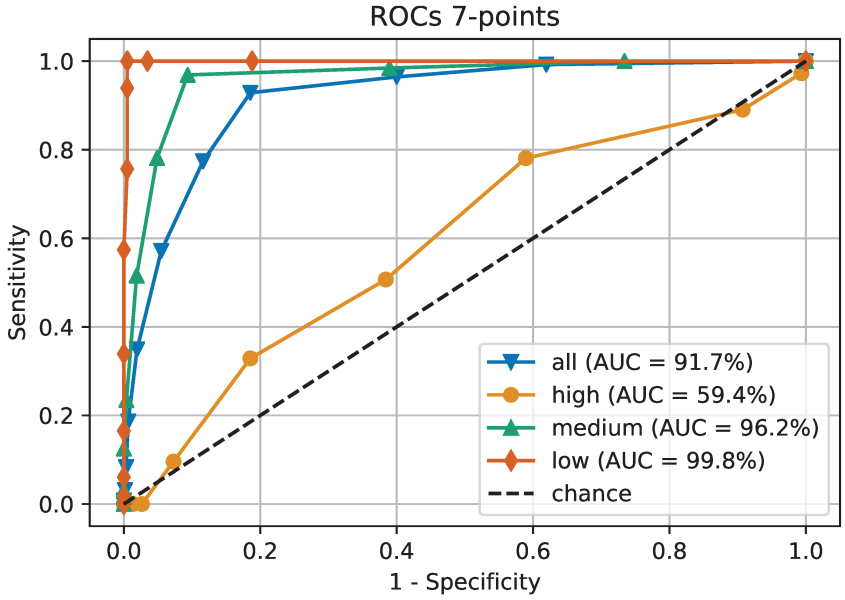


Figure 19 – Performance of the 7-point checklist algorithm on the Atlas dataset. It shows a huge gap to the performance of dermatologists evaluated in 100 random dermoscopic samples from the ISIC Archive, which is 67% Brinker et al. [2019]. The results for 7-point checklist applied on Atlas is optimistic considering the dataset's bias towards its educational aspects. Image reproduced from Bissoto et al. [2019].

299×299 to fit the input size of Inception-v4. We employ train and test data augmentation, with strategies described by Perez et al. [2018]. For all experiments, we report the Area Under the ROC Curve (AUC). Since both our datasets are relatively small, we choose not to use a validation set, using the weights after the 60th epoch for test evaluation[10].

We performed the following three experimental designs:

- **Destructing Atlas-dataset**: We employ the Atlas dataset with our disruptive actions for both training and testing the network in the destruction of information approach. We use 10 splits that we keep the same throughout all sets of images (*Traditional, Only Skin, Bounding Box, Bounding Box 70%*) to make comparisons fair. To compose each training split, we randomly select 70% of the images of each diagnostic difficulty present in the Atlas dataset (low, medium and high). We compose the corresponding test split using the 30% that is left. Following this procedure, we reduce the possibility of biasing our results with a split that is especially good for a given set of images. Since the training and test sets come from the same data distribution (same dataset), we expect these results to be optimistic, and that motivates our two next designs.

- **Destructing ISIC-dataset**: We also apply the destruction of information approach to the ISIC dataset. We do that to confirm the behavior verified in Atlas in a

[10] All our source code is readily available on `https://github.com/alceubissoto/deconstructing-bias-skin-lesion`.

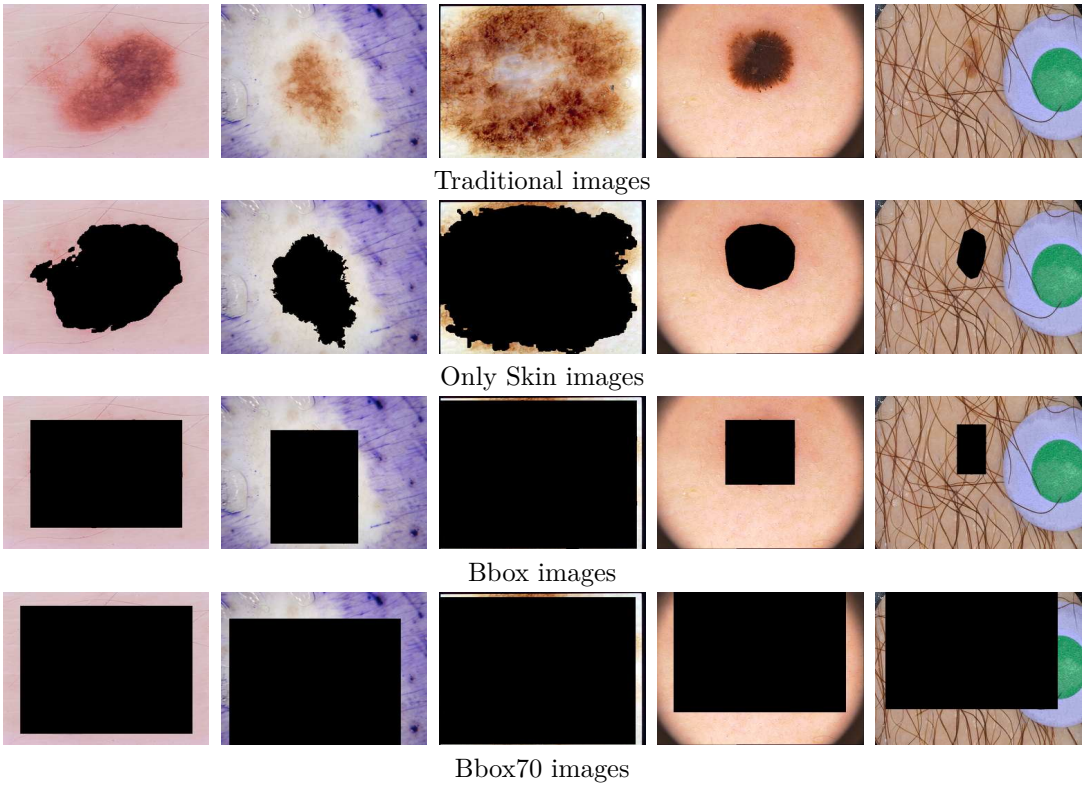


Figure 20 – Samples from each of our disrupted datasets. We gradually remove cogent information, until there is no information left to apply any aspect of medical score algorithms [Argenziano et al., 1998; Friedman et al., 1985]. Next, we use those sets to evaluate if the network can still learn patterns with the information left to correctly classify skin lesions. Best seen in digital format. Image reproduced from Bissoto et al. [2019].

more generic dataset, with fewer effects of human bias. We apply the same 10 split generation procedure we described for this experiment, except for the diagnostic difficulty stratification (the information is not present for the ISIC dataset).

- **Destructing Cross-dataset**: We increase the difficulty by experimenting with a cross-dataset fashion. We train with all $2,594$ samples from the ISIC dataset and evaluate on the complete 872 images set from Atlas. The differences between the statistics between those two datasets make this task harder, and better reflect a real-world setting [Torralba and Efros, 2011]. We repeat that experiment 10 times, for statistical significance.

4.8.2 Results and Analyses

Figures 21 and 22 show the network's performance for the different sets.

High difficulty lesions classification seem to be a very hard and specific task to the network, as it is for dermatologists. It could not learn clinical patterns properly with the training set, and destroying information do not influence the results. We understand that the network is probably exploiting image acquisition artifacts and dataset bias.

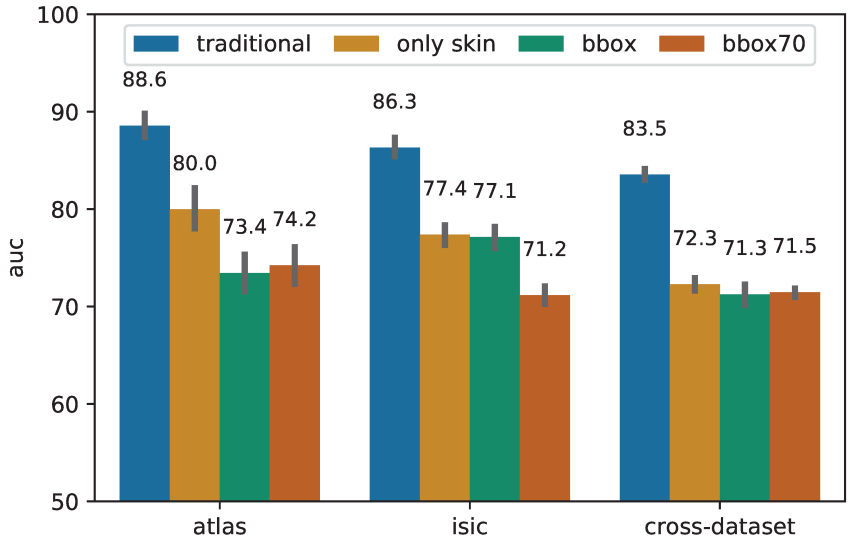


Figure 21 – Models' performance over the disturbed datasets. We first remove all the pixel colors inside the lesion (*only skin*), proceeding to remove border information (*bbox*), and finally, removing the size (diameter) of the lesion (*bbox70*). Surprisingly, even when we destruct all clinical-meaningful information, the network finds a way to learn to classify skin lesion images much better than chance. Image reproduced from Bissoto et al. [2019].

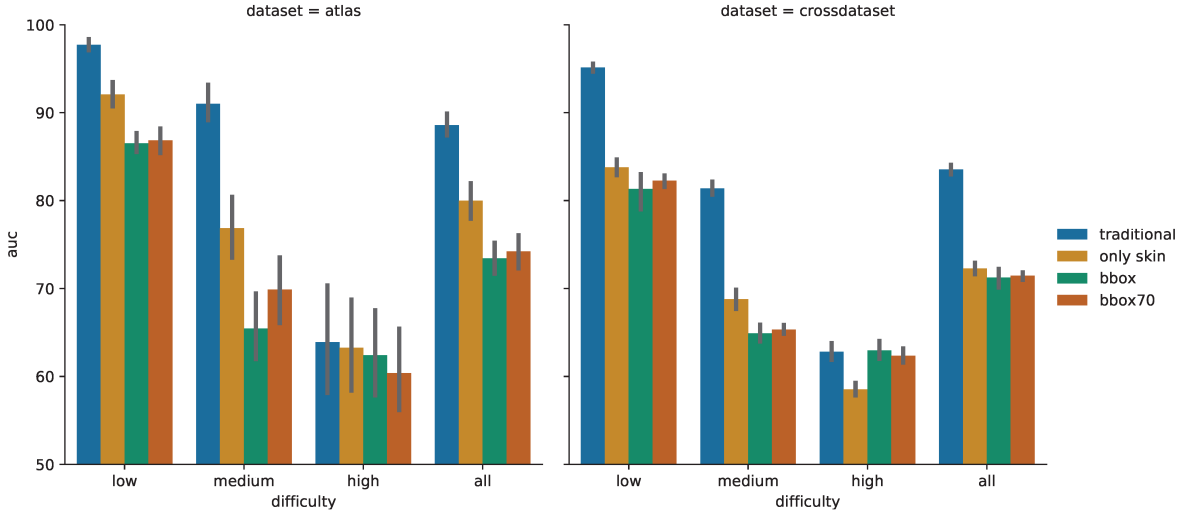


Figure 22 – We show the differences over the disturbed datasets, stratifying the performance into the different diagnostic difficulties. High difficulty diagnostic present resilience to the removal of cogent information. Despite not presenting as high numbers as the other difficulties, they are still much better than chance, revealing the patterns learned are not clinical. Other difficulties are more affected by the disturbances, but the overall result for *bbox*, and even *bbox70*, shockingly surpasses melanoma classification benchmark [Brinker et al., 2019] of 67%. This result suggests that dataset bias inflates our model's results. Image reproduced from Bissoto et al. [2019].

When experimenting in a cross-dataset fashion, the performance drops as expected, because of the differences between the statistics of Atlas and ISIC. The behavior of the network is similar in all experiments, and the following analysis can be generalized.

Traditional has the best overall performance, as expected. The network results follow the annotation of difficulty to diagnose by dermatologists. The results start to drop in *Only Skin*, where we start to deconstruct the information. When we remove the pixel information inside the lesion, we are removing all the information about dermoscopic attributes. The only clinically-meaningful information present is the border of the lesion, that could be used to verify its symmetry and irregularity, and skin features, such as vascularization.

When we remove the information of the borders, on *Bbox*, the performance lower, even more, revealing that we removed an essential feature for classification. An explanation, referring to medical algorithms like ABCD [Friedman et al., 1985], is that the diameter of the box contains the information on the size of the lesion, which is also relevant information when diagnosing skin lesions.

At *Bbox*70, we remove 70% of all pixels in the image and all medical relevant features that could aid the classification. Still, surprisingly, the network can make sense of visual features to make decisions that are much better than chance. There is a pattern within the available pixels that contain information that leads to the correct label. This is shocking. The numbers achieved by the network at this point even surpass the AUC achieved by dermatologists on the melanoma classification benchmark. As sanity check, we performed an experiment hiding all image information, feeding the network (for training and testing) only zero-filled images. We achieved an AUC of 50%, which is expected since AUC is insensitive to class balance.

We believe that dataset bias is the culprit for inflating the network's performance in our destructive experiments, introducing artifacts [Mishra and Celebi, 2016] that undesirably can deviate the network's attention from more critical features. We also verify that bias is not only present in the smaller educational purpose Atlas dataset, but also the most diversified ISIC dataset. Even performing the experiments in a cross-dataset fashion (the network is trained on ISIC, and tested on Atlas), the unnatural behavior persists, attesting to the fact that these two datasets may also share the same bias. We will address the exact causes and artifacts in future works.

Another possibility is that there is meaningful information at the borders of the images (parts that were not affected by the destruction procedures). This is unlikely because according to medical algorithms [Argenziano et al., 1998; Friedman et al., 1985; Abbasi et al., 2004], there is no information left to account.

Summarizing, the main contributions of these experiments are:

- We provide a discussion to raise awareness of bias in the automated skin lesion analysis community to improve the next generation of solutions for classifying skin lesions in the real world.

- We perform single- (training and testing on the same dataset) and cross-dataset (training on ISIC and testing on Atlas) experiments, and find that in both cases, the networks are able to maintain a surprising amount of accuracy, even after almost all cogent information has been destroyed.

4.9 Conclusion

This Chapter showed all the hypotheses and experiments conducted throughout the research. We were always guided by the desire to raise the quality and reliability of current techniques to allow their usage in real scenarios. Therefore, our contributions — much more than just reaching new state-of-the-art results — discuss methodological phenomena delaying the continuous improvement of existing methods.

Starting from a more in-depth analysis of the use of transfer learning — a practice consecrated in the literature of image classification and automated melanoma screening — we were able to elucidate the real contribution of this technique to the central problem, also listing a series of new perspectives that may direct future work, such as the use of other medical datasets to refine melanoma models.

We also focus on the experimental design of image classification techniques, proposing a robust framework specially designed for automatic screening of skin lesions. The main finding is that the size of training data has disproportionate influence, explaining almost half the variation in performance. Of the other factors, test data augmentation and input resolution are the most helpful. Deeper models — when combined, with extra data — also help. We show that the costly full factorial design, or the unreliable sequential optimization, are not the only options: ensembles of models provide reliable results with limited resources. Those results are relevant for future work on automated melanoma screening since they indicate which aspects of deep learning should be scrutinized to deliver better results.

In addition to more reliable experimental designs, we also analyze the quality of the datasets used in related research, showing the problems that the biases of such data causes, and raising many concerns about data that should be taken into account in future work.

Finally, besides participating in international competitions for skin lesion analysis through images, we distribute our codes for use by the community[11], in order to increase

[11] All source code of this work is available at: https://sites.google.com/site/robustmelanomascreening

the reproducibility of related literature.

In the next Chapter, we analyze our contributions to the current art and discuss future works.

5 Conclusion

In this final chapter, we summarize the major contributions and findings of this work, and discuss exciting directions for future works — discoveries and reflections that we could not address in this thesis due to limitation of time and scope.

5.1 Contributions

In this work we studied the current art of automated melanoma screening in an interdisciplinary way. We aimed to identify the main aspects of this research field, with a focus on deployment to the real world. We saw that computational, medical, and legal requirements must be addressed to enable such deployment.

Although the field has experienced a sharp advance in recent years, especially in what concerns Computer Sciences, there was still plenty to be done: in this work we proposed, evaluated, and delivered several experiments towards developing robust deep learning approaches for melanoma screening. We also surveyed the legal requirements to evaluate artificial intelligence solutions for healthcare. We saw that methodological gaps of the current literature impose important challenges on the rigorous evaluation of existing solutions, which may, in turn, affect how new solutions are evaluated by regulatory agencies, and affect the whole chain of deployment of solutions on the market. That is concerning, considering that software and hardware devices for assisted analysis of skin lesions are already being commercialized.

Finally, in the medical area, we investigated an interdisciplinary agenda of collaboration with physicians, concluding that those partnerships are crucial for speeding up the improvement of current art and creation of solutions aligned to real needs.

Our main contributions are:

- an interdisciplinary meta-analysis of current literature;
- a series of machine learning methodological guidelines to develop automated melanoma systems better;
- new perspectives for future works, enriching the discussions of current art.

5.2 Future Work

In this section, we discuss open challenges and future perspectives to develop automated melanoma screening for real-world scenarios. From a practical point of view,

recent work shows that automatic methods can overcome human rates of recognition of malignant skin lesions [Esteva et al., 2017; Haenssle et al., 2018; Tschandl et al., 2019]. However, these experiments are conducted under controlled conditions that do not necessarily reflect clinical reality.

Our work has shown that the degradation of the accuracy rates of machine learning models from conception to implantation is also a phenomenon observed in the melanoma screening literature. Mitigating this challenge is a critical step to allow the use of such systems in clinical situations. We have both demonstrated feasible steps that can reduce the excessive optimism on how models are evaluated (Section 4.6), and showcased how current datasets may give biased hits to AI models that are not necessarily biomedical relevant (Section 4.8).

From a broader perspective, to promote real-world screening solutions, we believe that future research must address the following six aspects:

- **know which system to develop**: Artificial Intelligence is gaining more attention for medical purposes. Researchers should deliberate about the implications of having CAD software of that nature deployed in the public or private health systems: are patients ready to receive a diagnosis from a machine? We argue that a *referability* framework is 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 the *hard* cases (for which the model has low confidence), might be more achievable in the short term than attempting to have high confidence for all cases;

- **broaden the availability of data for research**: our experiments have shown the importance of data for models based on deep learning, so continuing efforts to capture and disseminate them must be a continuous effort. Other possibilities would be working with the already available data, augmenting the datasets artificially. For examples, with typical data augmentation approaches, or with another kind of augmentation, like the ones inspired in biological aspects of the disease [Vasconcelos and Vasconcelos, 2017]. Another approach that should be investigated is the usage of GANs to generate artificial images of skin lesions that are so convincing as to deceive physicians about their artificiality [Bissoto et al., 2018b];

- **incorporate clinical data from patients**: we performed very few experiments on the benefit of using patient data in model construction. The literature itself is inconclusive about whether that information helps to deliver the diagnosis or not. However, clinical data is something that future research should investigate adequately. Dermatologists use this kind of information — e.g., if the lesion is growing, if it itches, if it bleeds, if it hurts, its location and patient's age and sex — to better

diagnose skin lesions. So, why they are not fully incorporated in current methods for automated screening?

- **be aware of medical procedures**: the medical algorithms to detect malignant skin lesions help improving human accuracy rates. However, if future research continue trying to implement such as they are, we believe that the predictive power of the models will be limited, because machines "think" different as humans; instead, we argue that trying to *understand* the rationale behind such algorithms would be more helpful to incorporate medical knowledge on modern approaches;

- **increase the variability of image sources in reference datasets**: rather than acquiring new images, it is interesting that public databases contain data from different sources to broaden the representation of the population. The biggest problem when implementing a real-time automatic classification system is that the system will have access to data many times different from the ones used in its construction. Decreasing this gap is a way to avoid the natural degradation of the models;

- **investigate standardized forms for skin lesion imaging**: variability of data sources and acquisition processes make models more generalizable and robust for real-life use. However, stipulating a standardized data acquisition can be beneficial, because the machine no longer has to learn issues related to lighting, resolution, brightness, and can focus on more relevant nuances such as skin patterns, prototypes, and skin aging.

From the computing point of view, we believe that past models (global descriptors, BoVWs) can no longer compete with DNNs, even in the context of small-dataset medical applications. If some model can beat DNNs, it is a model from the future, not from the past. In our group, our current research efforts on automated diagnosis rely on DNNs, for example, exploring tuning approaches, deeper models, and assembling of different techniques. Since 2015 the literature has presented solutions based on deep learning, generating better results. We believe that this path is no longer back, so future works should surely invest in this kind of approach.

In addition to lesion classification aiming at diagnosis, automated screening involves other relevant tasks: lesion detection, segmentation, registration, and tracking. For many of those tasks the main aim is building/analysing a time series of lesion evolution. Indeed, as techniques like full-body skin scan increase in popularity, automating those tasks may become critical — but currently there is essentially no publicly available dataset allowing researchers to work on lesion tracking or evolution.

We believe it would benefit everyone if researchers were more aware of regulatory processes, and had them in mind when designing and conducting their experiments. Such

knowledge should ideally inform everything from the quality of the code in current prototypes, until future clinical trials that hopefully will happen when the field matures. Although not all published works are interested in launching market products for automated screening, quality concerns would benefit the entire field: including colleagues who are attempting to reproduce others' works. Researchers should also be aware that their works, even when not intended for clinical settings, set a "background standard" of how techniques are perceived/evaluated by regulatory agencies.

An exciting new frontier for automated skin analysis is the arrival of new types of data, including new types of images. Newer imaging technologies such as infrared imaging, multispectral imaging, and confocal microscopy, have recently come to the forefront in providing the potential for greater diagnostic accuracy. Again, the limitating factor is the lack of publicly available datasets: we hope to see new developments in those areas, that allow collaboration in the acquisition of these types of data.

Advances in the literature have already brought significant changes on melanoma screening. Artificial Intelligence for automated lesion analysis is already here. We still fall short of deploying accurate, robust systems in real-world clinical settings, but the directions above indicate many possibilities to make that possible. We are excited to see what the future holds.

5.3 Publications

We highlight the following publications, results of the Ph.D.:

Journal Papers:

- Valle, E., Fornaciali, M., Menegola, A., Tavares, J., Bittencourt, F. V., Li, L. T., & Avila, S. (2017). Data, Depth, and Design: Learning Reliable Models for Melanoma Screening. (in press at Neurocomputing)

Conference Papers:

- Bissoto, A., Fornaciali, M., Valle, E., & Avila, S. (2019). (De) Constructing Bias on Skin Lesion Datasets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 0-0).
- Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., & Valle, E. (2017, April). Knowledge transfer for melanoma screening with deep learning. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) (pp. 297-300). IEEE.

Preprints:

- Fornaciali, M., Carvalho, M., Bittencourt, F. V., Avila, S., & Valle, E. (2016). Towards automated melanoma screening: Proper computer vision & reliable results. arXiv preprint arXiv:1604.04024.

Technical Reports:

- Bissoto, A., Perez, F., Ribeiro, V., Fornaciali, M., Avila, S., & Valle, E. (2018). Deep-Learning Ensembles for Skin-Lesion Segmentation, Analysis, Classification: RECOD Titans at ISIC Challenge 2018. arXiv preprint arXiv:1808.08480.
- Menegola, A., Tavares, J., Fornaciali, M., Li, L. T., Avila, S., & Valle, E. (2017). RECOD titans at ISIC challenge 2017. arXiv preprint arXiv:1703.04819.

5.4 Achievements

This research also received the following **prizes and distinctions**:

- *Google Research Awards for Latin America* (2016 and 2017)
- 1st place on "melanoma classification task" @ ISIC Challenge 2017
- Honorable mention of a post presentation of our first conference paper Menegola et al. [2017a] @ "*12ª Conferência Brasileira sobre Melanoma*", São Paulo/SP (2017)
- 5th place (4th team) on "lesion diagnosis task — challenge data only" @ ISIC Challenge 2018
- 2nd place in the "Best Paper Awards" of the conference paper Bissoto et al. [2018b] @ "*2nd International Educational Symposium of the Melanoma World Society (MWS)*", Rio de Janeiro/RJ (2018)

Bibliography

International Skin Imaging Collaboration: Melanoma Project. `https://isic-archive.com`. Cited 2 times in pages 32 and 35.

Abbas, Q., Emre Celebi, M., Garcia, I. F., and Ahmad, W. Melanoma recognition framework based on expert definition of abcd for dermoscopic images. *Skin Research and Technology*, 19(1), 2013. Cited in page 35.

Abbasi, N. R., Shaw, H. M., Rigel, D. S., Friedman, R. J., McCarthy, W. H., Osman, I., Kopf, A. W., and Polsky, D. Early diagnosis of cutaneous melanoma: revisiting the abcd criteria. *Jama*, 292(22):2771–2776, 2004. Cited in page 83.

Ali, A.-R. A. and Deserno, T. M. A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data. In *Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*, volume 8318, page 83181I. International Society for Optics and Photonics, 2012. Cited in page 61.

Allison, D. B., Brown, A. W., George, B. J., and Kaiser, K. A. Reproducibility: A tragedy of errors. *Nature*, 530(7588):27–29, 2016. Cited in page 48.

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., and Delfino, M. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of dermatology*, 134(12):1563–1570, 1998. Cited 5 times in pages 49, 78, 79, 81, and 83.

Argenziano, G., Soyer, H. P., De Giorgi, V., Piccolo, D., Carli, P., Delfino, M., et al. Dermoscopy: a tutorial. *EDRA, Medical Publishing & New Media*, 2002. Cited 2 times in pages 16 and 54.

Argenziano, G., Catricalà, C., Ardigo, M., Buccini, P., De Simone, P., Eibenschutz, L., Ferrari, A., Mariani, G., Silipo, V., Sperduti, I., et al. Seven-point checklist of dermoscopy revisited. *British Journal of Dermatology*, 164(4):785–790, 2011. Cited in page 49.

Avila, S., Thome, N., Cord, M., Valle, E., and de A. Araújo, A. Pooling in image representation: the visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465, 2013. Cited in page 29.

Ballerini, L., Fisher, R. B., Aldridge, B., and Rees, J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. 2013. Cited in page 55.

Barata, C., Celebi, M. E., and Marques, J. S. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE journal of biomedical and health informatics*, 23 (3):1096–1109, 2018. Cited in page 28.

Bi, L., Kim, J., Ahn, E., and Feng, D. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. *arXiv preprint arXiv:1703.04197*, 2017. Cited 2 times in pages 34 and 35.

Bissoto, A., Perez, F., Ribeiro, V., Fornaciali, M., Avila, S., and Valle, E. Deep-learning ensembles for skin-lesion segmentation, analysis, classification: Recod titans at isic challenge 2018. *arXiv preprint arXiv:1808.08480*, 2018a. Cited in page 75.

Bissoto, A., Perez, F., Valle, E., and Avila, S. Skin lesion synthesis with generative adversarial networks. In *ISIC Skin Image Analysis Workshop and Challenge @ MICCAI 2018*, 2018b. Cited 4 times in pages 21, 76, 87, and 90.

Bissoto, A., Fornaciali, M., Valle, E., and Avila, S. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. Cited 4 times in pages 78, 80, 81, and 82.

Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A. H., and von Kalle, C. Skin cancer classification using convolutional neural networks: systematic review. *Journal of medical Internet research*, 20(10):e11936, 2018. Cited in page 23.

Brinker, T. J., Hekler, A., Hauschild, A., Berking, C., Schilling, B., Enk, A. H., Haferkamp, S., Karoglan, A., von Kalle, C., Weichenthal, M., et al. Comparing artificial intelligence algorithms to 157 german dermatologists: the melanoma classification benchmark. *European Journal of Cancer*, 111:30–37, 2019. Cited 3 times in pages 79, 80, and 82.

Carli, P., Quercioli, E., Sestini, S., Stante, M., Ricci, L., Brunasso, G., and De Giorgi, V. Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. *British Journal of Dermatology*, 148(5):981–984, 2003. Cited 2 times in pages 39 and 49.

Carvalho, M. Transfer schemes for deep learning in image classification. Master's thesis, University of Campinas, 2015. Cited 2 times in pages 18 and 53.

Celebi, M. E., Wen, Q., Iyatomi, H., Shimizu, K., Zhou, H., and Schaefer, G. A state-of-the-art survey on lesion border detection in dermoscopy images, 2015. Cited in page 25.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. Cited in page 58.

Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., and Smith, J. R. Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In *Machine Learning in Medical Imaging*, pages 118–126. Springer, 2015. Cited 5 times in pages 32, 33, 34, 35, and 67.

Codella, N. C., Nguyen, Q.-B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., and Smith, J. R. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4):5:1–5:15, 2017. Cited 2 times in pages 34 and 35.

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 168–172, 2018. Cited 3 times in pages 32, 35, and 56.

Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019. Cited 2 times in pages 32 and 56.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. Cited 2 times in pages 28 and 32.

Day, G. R. and Barbour, R. H. Automated melanoma diagnosis: where are we at? *Skin Research and Technology*, 6(1):1–5, 2000. Cited 4 times in pages 22, 23, 30, and 31.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. Cited in page 57.

Devesa, F. R. S. Medical software certification processes in europe, usa and brazil. Master's thesis, 2014. Cited 2 times in pages 40 and 41.

DeVries, T. and Ramachandram, D. Skin lesion classification using deep multi-scale convolutional neural networks. *arXiv preprint arXiv:1703.01402*, 2017. Cited 2 times in pages 34 and 35.

Díaz, I. G. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. *arXiv preprint arXiv:1703.01976*, 2017. Cited 2 times in pages 34 and 35.

Eakins, J. and Graham, M. Content-based image retrieval. 1999. Cited in page 29.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017. Cited 8 times in pages 17, 18, 33, 34, 35, 39, 41, and 87.

Fernandes, S. L., Chakraborty, B., Gurupur, V. P., Prabhu, G., et al. Early skin cancer detection using computer aided diagnosis techniques. *Journal of Integrated Design and Process Science*, 20(1):33–43, 2016. Cited in page 22.

Food, U. and Administration, D. Fda permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. *FDA News Release*, 2018. Cited 2 times in pages 17 and 44.

Fornaciali, M. Towards robust melanoma screening: A case for enhanced mid-level features. Master's thesis, University of Campinas, 2015. Cited 4 times in pages 16, 24, 29, and 53.

Fornaciali, M., Avila, S., Carvalho, M., and Valle, E. Statistical learning approach for robust melanoma screening. In *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 319–326, 2014. Cited 2 times in pages 18 and 53.

Fornaciali, M., Carvalho, M., Bittencourt, F. V., Avila, S., and Valle, E. Towards automated melanoma screening: Proper computer vision & reliable results. *arXiv preprint arXiv:1604.04024*, 2016. Cited 6 times in pages 23, 24, 27, 48, 53, and 77.

Friedman, R. J., Rigel, D. S., and Kopf, A. W. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians*, 35(3):130–151, 1985. Cited 4 times in pages 31, 49, 81, and 83.

Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980. Cited in page 26.

Ge, Z., Demyanov, S., Bozorgtabar, B., Abedini, M., Chakravorty, R., Bowling, A., and Garnavi, R. Exploiting local and generic features for accurate skin lesions classification using clinical and dermoscopy imaging. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 986–990, 2017a. Cited 2 times in pages 34 and 35.

Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., and Garnavi, R. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 250–258. Springer, 2017b. Cited in page 35.

Gutman, D., Codella, N., Celebi, E., Helba, B., Marchetti, M., Mishra, N., and Halpern, A. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. Cited 3 times in pages 32, 35, and 56.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018. Cited 4 times in pages 17, 18, 33, and 87.

Hand, M., Chien, A., and Grossman, D. Screening and non-invasive evaluative devices for melanoma detection: A comparison of commercially available devices and dermoscopic evaluation. *Journal of Clinical Dermatology and Therapy*, 2015. Cited in page 17.

Harangi, B. Skin lesion detection based on an ensemble of deep convolutional neural network. *arXiv preprint arXiv:1705.03360*, 2017. Cited 2 times in pages 34 and 35.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. Cited 2 times in pages 62 and 76.

Hinton, G., Osindero, S., and Teh, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006. Cited in page 26.

Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. Densely connected convolutional networks. In *IEEE CVPR*, 2017. Cited in page 76.

IEC 62304:2006. Medical device software – Software life cycle processes. Standard, International Organization for Standardization, Geneva, CH, 2006. Cited 2 times in pages 41 and 43.

ISO 13485:2003. Medical devices – Quality management systems – Requirements for regulatory purposes. Standard, International Organization for Standardization, Geneva, CH, 2003. Cited 2 times in pages 40 and 41.

ISO 9001:2008. Quality management systems – Requirements. Standard, International Organization for Standardization, Geneva, CH, 2008. Cited in page 41.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, 2010. Cited in page 29.

Jia, X. and Shen, L. Skin lesion classification using class activation map. *arXiv preprint arXiv:1703.01053*, 2017. Cited 2 times in pages 34 and 35.

Kawahara, J., BenTaieb, A., and Hamarneh, G. Deep features to classify skin lesions. In *IEEE International Symposium on Biomedical Imaging*, pages 1397–1400, 2016. Cited 2 times in pages 34 and 35.

Kharazmi, P., Zheng, J., Lui, H., Wang, Z. J., and Lee, T. K. A computer-aided decision support system for detection and localization of cutaneous vasculature in dermoscopy images via deep feature learning. *Journal of medical systems*, 42(2):33, 2018. Cited in page 39.

Korotkov, K. and Garcia, R. Computerized analysis of pigmented skin lesions: a review. *Artificial intelligence in medicine*, 56(2):69–90, 2012. Cited 8 times in pages 17, 20, 21, 22, 23, 26, 30, and 31.

Krizhevsky, A., Sutskever, I., and Hinton, G. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pages 1–9, 2012. Cited 2 times in pages 26 and 31.

Kumar, M. and Singh, K. M. Content based medical image retrieval system using dwt and lbp for ear images. *International Science Press*, 9(40):353–358, 2016. Cited in page 30.

Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006. Cited in page 29.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 0028-0836. Cited 2 times in pages 17 and 27.

Li, P., Lu, X., and Wang, Q. From dictionary of visual words to subspaces: Locality-constrained affine subspace coding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2348–2357, 2015. Cited in page 29.

Lisboa, P. J. and Taktak, A. F. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, 19(4):408–415, 2006. Cited 2 times in pages 23 and 24.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical

image analysis. *Medical image analysis*, 42:60–88, 2017. Cited 2 times in pages 17 and 26.

Lopez, A. R., Giro-i Nieto, X., Burdick, J., and Marques, O. Skin lesion classification from dermoscopic images using deep learning techniques. In *IEEE International Conference on Biomedical Engineering (BioMed)*, pages 49–54, 2017. Cited 2 times in pages 34 and 35.

Madooei, A. and Drew, M. S. Incorporating colour information for computer-aided diagnosis of melanoma from dermoscopy images: a retrospective survey and critical analysis. *International journal of biomedical imaging*, 2016, 2016. Cited in page 25.

Malvehy, J., Puig, S., Argenziano, G., Marghoob, A. A., and Soyer, H. P. Dermoscopy report: proposal for standardization: results of a consensus meeting of the international dermoscopy society. *Journal of the American Academy of Dermatology*, 57(1):84–95, 2007. Cited in page 32.

Mar, V. J., Scolyer, R. A., and Long, G. V. Computer-assisted diagnosis for skin cancer: have we been outsmarted? *The Lancet*, 389(10083):1962–1964, 2017. Cited in page 39.

Marchetti, M. A., Codella, N. C., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., Mishra, N., Carrera, C., Celebi, M. E., DeFazio, J. L., et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology*, 78(2):270–277, 2018. Cited 2 times in pages 39 and 41.

Masood, A. and Ali Al-Jumaily, A. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *International Journal of Biomedical Imaging*, 2013:22, 2013. Cited 2 times in pages 22 and 48.

Masood, A., Al-Jumaily, A., and Anam, K. Self-supervised learning model for skin cancer diagnosis. In *International IEEE/EMBS Conference on Neural Engineering*, pages 1012–1015, 2015. Cited 2 times in pages 32 and 33.

Matsunaga, K., Hamada, A., Minagawa, A., and Koga, H. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*, 2017. Cited 2 times in pages 34 and 35.

McCulloch, W. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of Mathematical Biophysics*, 5(4):115–133, 1943. Cited in page 26.

Mendonca, T., Ferreira, P. M., Marques, J. S., Marcal, A. S., and Rozeira, J. Ph2 – a dermoscopic image database for research and benchmarking. In *IEEE Engineering in*

Medicine and Biology Society, pages 5437–5440, July 2013. Cited 2 times in pages 32 and 55.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., and Valle, E. Knowledge transfer for melanoma screening with deep learning. In *IEEE International Symposium on Biomedical Imaging*, pages 297–300, 2017a. Cited 9 times in pages 28, 34, 35, 57, 58, 59, 60, 71, and 90.

Menegola, A., Tavares, J., Fornaciali, M., Li, L. T., Avila, S., and Valle, E. RECOD Titans at ISIC Challenge 2017. *arXiv preprint arXiv:1703.04819*, 2017b. Cited 4 times in pages 34, 35, 61, and 65.

Menegola, A. et al. Deep learning in melanoma screening= aprendizado profundo em triagem de melanoma. 2018. Cited in page 25.

Menzies, S. W. Automated epiluminescence microscopy: human vs machine in the diagnosis of melanoma. *Archives of dermatology*, 135(12):1538–1540, 1999. Cited in page 39.

Mishra, N. K. and Celebi, M. E. An overview of melanoma detection in dermoscopy images using image processing and machine learning. *arXiv preprint arXiv:1601.07843*, 2016. Cited 2 times in pages 78 and 83.

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S. M. R., Jafari, M. H., Ward, K., and Najarian, K. Melanoma detection by analysis of clinical images using convolutional neural network. In *IEEE Engineering in Medicine and Biology Society*, pages 1373–1376, 2016. Cited 2 times in pages 34 and 35.

of Health, U. D. and Services, H. Software as a medical device (samd): Clinical evaluation. guidance for industry and food and drug administration staff. 2017. Cited in page 43.

Oliveira, R. B., Papa, J. P., Pereira, A. S., and Tavares, J. M. R. Computational methods for pigmented skin lesion classification in images: review and future trends. *Neural Computing and Applications*, pages 1–24, 2016. Cited in page 22.

Pathan, S., Prabhu, K. G., and Siddalingaswamy, P. Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—a review. *Biomedical Signal Processing and Control*, 39:237–262, 2018. Cited 2 times in pages 23 and 26.

Pehamberger, H., Steiner, A., and Wolff, K. In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions. *Journal of the American Academy of Dermatology*, 17(4):571–583, 1987. Cited in page 49.

Peng, R. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011. ISSN 0036-8075. Cited in page 48.

Perez, F., Vasconcelos, C., Avila, S., and Valle, E. Data augmentation for skin lesion analysis. In *ISIC Skin Image Analysis Workshop and Challenge @ MICCAI 2018*, 2018. Cited 4 times in pages 34, 76, 77, and 80.

Perez, F., Avila, S., and Valle, E. Solo or ensemble? choosing a cnn architecture for melanoma classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. Cited in page 34.

Perronnin, F., Sánchez, J., and Mensink, T. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010. Cited in page 29.

Picard, D. and Gosselin, P.-H. Improving image similarity with vectors of locally aggregated tensors. In *IEEE International Conference on Image Processing*, pages 669–672, 2011. Cited in page 29.

Pires, R., Avila, S., Wainer, J., Valle, E., Abramoff, M. D., and Rocha, A. A data-driven approach to referable diabetic retinopathy detection. *Artificial intelligence in medicine*, 96:93–106, 2019. Cited in page 17.

Premaladha, J. and Ravichandran, K. Asymmetry analysis of malignant melanoma using image processing: a survey. *Journal of Artificial Intelligence*, 7(2):45–53, 2014. Cited in page 25.

Premaladha, J. and Ravichandran, K. Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms. *Journal of Medical Systems*, 40(4):1–12, 2016. Cited in page 33.

Rajpara, S., Botello, A., Townend, J., and Ormerod, A. Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. *British Journal of Dermatology*, 161(3):591–604, 2009. Cited in page 39.

Ribeiro, V., Avila, S., and Valle, E. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv preprint arXiv:1906.02415*, 2019. Cited in page 68.

Rigel, D. S. Epidemiology of melanoma. In *Seminars in cutaneous medicine and surgery*, volume 29, pages 204–209. WB Saunders, 2010. Cited in page 16.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015. Cited in page 67.

Rosado, B., Menzies, S., Harbauer, A., Pehamberger, H., Wolff, K., Binder, M., and Kittler, H. Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis. *Archives of Dermatology*, 139(3):361–367, 2003. Cited in page 39.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. Cited 2 times in pages 78 and 79.

Sabbaghi, S., Aldeen, M., and Garnavi, R. A deep bag-of-features model for the classification of melanomas in dermoscopy images. In *IEEE Engineering in Medicine and Biology Society*, pages 1369–1372, 2016. Cited 2 times in pages 33 and 35.

Safran, T., Viezel-Mathieu, A., Corban, J., Kanevsky, A., Thibaudeau, S., and Kanevsky, J. Machine learning and melanoma: The future of screening. *Journal of the American Academy of Dermatology*, 2017. Cited in page 39.

Salzberg, S. L. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997. Cited in page 77.

Sandve, G., Nekrutenko, A., Taylor, J., and Hovig, E. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, 10 2013. Cited in page 48.

Sathiya, S. B., Kumar, S., and Prabin, A. A survey on recent computer-aided diagnosis of melanoma. In *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on*, pages 1387–1392. IEEE, 2014. Cited in page 22.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Cited in page 58.

Sinz, C., Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., Cabo, H., Gourhant, J.-Y., Kreusch, J., et al. Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *Journal of the American Academy of Dermatology*, 77(6):1100–1109, 2017. Cited in page 31.

Sivic, J. and Zisserman, A. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003. Cited 2 times in pages 28 and 32.

Soyer, H., Argenziano, G., Zalaudek, I., Corona, R., Sera, F., Talamini, R., Barbato, F., Baroni, A., Cicale, L., Di Stefani, A., et al. Three-point checklist of dermoscopy. *Dermatology*, 208(1):27–31, 2004. Cited in page 49.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017. Cited 3 times in pages 62, 76, and 79.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011. Cited 2 times in pages 78 and 81.

Tschandl, P., Rosendahl, C., and Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5: 180161, 2018. Cited in page 56.

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 2019. Cited 4 times in pages 17, 18, 33, and 87.

Tuong, W., Cheng, L. S., and Armstrong, A. W. Melanoma: epidemiology, diagnosis, treatment, and outcomes. *Dermatologic clinics*, 30(1):113–124, 2012. Cited in page 16.

Valle, E., Fornaciali, M., Menegola, A., Tavares, J., Bittencourt, F. V., Li, L. T., and Avila, S. Data, depth, and design: Learning reliable models for melanoma screening. *arXiv preprint arXiv:1711.00441*, 2017. Cited 9 times in pages 33, 35, 66, 67, 69, 72, 74, 75, and 77.

Vasconcelos, C. N. and Vasconcelos, B. N. Experiments using deep learning for dermoscopy image analysis. *Pattern Recognition Letters*, 2017. Cited 2 times in pages 34 and 87.

Voss, R. K., Woods, T. N., Cromwell, K. D., Nelson, K. C., and Cormier, J. N. improving outcomes in patients with melanoma: strategies to ensure an early diagnosis. *Patient related outcome measures*, 6:229, 2015. Cited in page 16.

Wighton, P., Lee, T., Lui, H., McLean, D., and Atkins, M. Generalizing common tasks in automated skin lesion diagnosis. *IEEE Transactions on Information Technology in Biomedicine*, 15(4):622–629, 2011. Cited in page 35.

Yang, X., Zeng, Z., Yeo, S. Y., Tan, C., Tey, H. L., and Su, Y. A novel multi-task deep learning model for skin lesion segmentation and classification. *arXiv preprint arXiv:1703.01025*, 2017. Cited 2 times in pages 34 and 35.

Yoshida, T., Celebi, M. E., Schaefer, G., and Iyatomi, H. Simple and effective preprocessing for automated melanoma discrimination based on cytological findings. In *IEEE International Conference on Big Data*, pages 3439–3442, 2016. Cited 3 times in pages 35, 36, and 39.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014. Cited in page 28.

Yu, L., Chen, H., Dou, Q., Qin, J., and Heng, P.-A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017. Cited 2 times in pages 34 and 35.