Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Vinícius de Paulo Souza Ribeiro

The Impact of Annotation Quality on Deep Learning for Skin Lesion Segmentation

O Impacto da Qualidade das Anotações na Aprendizagem Profunda para a Segmentação de Lesões de Pele

Campinas

2019

Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Vinícius de Paulo Souza Ribeiro

The Impact of Annotation Quality on Deep Learning for Skin Lesion Segmentation

O Impacto da Qualidade das Anotações na Aprendizagem Profunda para a Segmentação de Lesões de Pele

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

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

Supervisor: Prof. Dr. Eduardo Alves do Valle Junior

Este exemplar corresponde à versão final da tese defendida pelo aluno Vinicius de Paulo Souza Ribeiro, e orientada pelo Prof. Dr. Eduardo Alves do Valle Junior

Campinas

2019

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

R354i

Ribeiro, Vinícius de Paulo Souza, 1993-
The impact of annotation quality on deep learning for skin lesion segmentation / Vinícius de Paulo Souza Ribeiro. – Campinas, SP : [s.n.], 2019.

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

1. Aprendizado de máquina. 2. Aprendizagem supervisionada (Aprendizado do computador). 3. Segmentação de imagens médicas. 4. Peles - Câncer. 5. Concordancias. I. Valle Junior, Eduardo Alves do. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: O impacto da qualidade das anotações na aprendizagem profunda para a segmentação de lesões de pele
Palavras-chave em inglês:
Machine learning
Supervised learning (Machine learning)
Medical images segmentation
Skin - Câncer
Agreement
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Eduardo Alves do Valle Junior [Orientador]
Roberto de Alencar Lotufo
Thiago Vallin Spina
Data de defesa: 05-08-2019
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-5897-5765
- Currículo Lattes do autor: http://lattes.cnpq.br/4867885877970077

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidato: Vinícius de Paulo Souza Ribeiro RA: 118906

Data da Defesa: 5 de agosto de 2019

Título da Tese: The Impact of Annotation Quality on Deep Learning for Skin Lesion Segmentation (O Impacto da Qualidade das Anotações na Aprendizagem Profunda para a Segmentação de Lesões de Pele).

Prof. Dr. Eduardo Alves do Valle Junior (Presidente)
Prof. Dr. Roberto de Alencar Lotufo
Dr. Thiago Vallin Spina

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 above all, I would like to thank my supervisor, Prof. Dr. Eduardo Valle. Prof. Valle is an example to follow both as a professional and as a human being, and a reference for me as a scientist. I had the pleasure to join his team in the mid of 2014 as an undergraduate student, and then I realized how important it is to have a supervisor you can trust and one that trusts you. I knew right from the beginning, we would have a fruitful relationship, and I have no regrets about choosing him to supervise my M. Sc. studies. Prof. Valle is reliable, ethical, and has a brilliant mind. He is probably one of the most promising scientists in Brazil, and I cannot see my studies being as successful as it was if it was not for his supervision. I am very proud of being his student for all these years.

I would also like to thank all the members of the Learning Titans research group. Profa. Dra. Sandra Avila, Michel Fornaciali, Eduardo Seiti, Fábio Perez, Alceu Bissoto and all the other people that joined our group along the years. All of them contributed to my research and had a material impact on my life, and many of them became close friends of mine. They not only brought essential ideas to the projects but also helped and supported me when things were not going well. This group is making a massive breakthrough in our research field, and we can expect to see great things coming in the future. I want to give a special thanks to Profa. Avila, which is an outlier, a very patient person, and a brilliant scientist. Even though I was not her direct student, she was always there to help and advice me.

My research would not be possible without the infrastructure provided by RECOD — Reasoning for Complex Data, our lab at the Institute of Computing (IC) at Unicamp, which provided the computational power to run our experiments. I want to thank the School of Electrical and Computer Engineering (FEEC) at Unicamp and all its professors, which contributed in some way to form me as a professional, a scientist and a human. I am proud of carrying Unicamp's name in my Curriculum Vitae. I hope one day I can return to society at least a fraction of what this incredible University provided me. In times of trouble and continuous threat to science, we must stay together and defend our institutions. Unicamp thought me to stay on the bright side, and fight for freedom, democracy, and science.

I want to give a special thanks to the people that stayed by my side for all my life and never left me alone, my parents, Wanius Ribeiro and Raquel de Souza Ribeiro, and my sister, Nicole de Souza Ribeiro. Having a family like the one I have is a privilege, and I will never be able to fully express how much I am grateful to have them by my side and how much I love them. They

are my basis, they built my character and thought me all the most important things a person must learn. They thought me to be a decent person, to love and to respect. Most importantly, they thought me to never give up on my values and dreams, and always fight for what is right. We stayed together during the most challenging times, and I could not be more proud of having them with me. This accomplishment would have no meaning without my family.

Last but not least, I would like to thank all my friends and coworkers. It would be unfeasible for me to list all of their names. Each of them knows the impact they have in my life and how important they are to me. I want to thank all my friends and relatives for friendship and continuous support. Mainly, I would like to name a few people that were closer to me during my studies and my life: Daniel Melges, Edgar Berg, Paulo Azevedo, and Marina Senese. Even when things were hard, and we could not be as close as I wanted, they were always there for me, and I am delighted to have their company. During my M. Sc. studies, I worked in many places and with many people. I am thankful to all of the companies I passed and to all the bosses I had, who were very positive about investing in me and making it feasible for me to work and study concurrently. They were all essential in my life and all the people I worked with made me a better engineer and a better scientist. Mainly, I am thankful to the current company I work, Nexa Digital, and the French Embassy at Brazil for providing the means and supporting me during my stay at France for the Summer School on Data Science for Document Analysis and Understanding in July 2019.

When the facts change, I change my mind. What do you do, Sir?

John Maynard Keynes

Abstract

Every year, the National Institute of Cancer, in Brazil, registers more than 150 000 new cases of skin cancer, making it a real issue in the country's public health system. Skin cancer evolves in different manners, the most common is the basal cell carcinoma, but melanoma is the most dangerous, with the highest mortality rate. The probability of cure decreases with the matureness of the disease. In this scenario, automatic methods for skin lesion triage is hope for boosting early detection and increasing the life expectancy of cancer patients. In this study, we address one of the main subjects of the skin cancer detection pipeline: skin lesion segmentation. The task itself is challenging from the computer vision perspective. Public data sets are not as large as for other image domains, and the annotations are not optimal. These problems have a real impact on the model's performance and capability to generalize. Along with our work, we aim to tackle the second issue, the quality of image ground truths. We analyze the inter-annotator agreement statistics inside the most popular skin lesion dataset public available and draw some conclusions about the available annotations. Then, we propose a series of conditioning to apply in the training data to evaluate how they improve the agreement between different specialists. Finally, we analyze how the conditionings affect the training and evaluation of deep neural networks for the skin lesion segmentation task. Our conclusions show that the low inter-annotator agreement available in the ISIC Archive dataset has a meaningful impact in the performance of trained models and taking the disagreement into account can indeed improve the generalization capability of the networks.

Resumo

Todos os anos, o Instituto Nacional do Câncer, no Brasil, registra mais de 150 000 novos casos de câncer de pele, configurando um problema real no sistema de saúde pública do país. O câncer de pele se desenvolve de maneiras diferentes, sendo o melanoma o mais perigoso, com a maior taxa de mortalidade. As chances de cura diminuem com o avanço da doença. Nesse cenário, métodos automáticos de triagem de lesões de pele abrem uma perspectiva para uma detecção mais precoce da doença, e um melhor prognóstico para os pacientes de câncer. Nesse estudo, nós endereçamos uma das principais tarefas do pipeline de deteção de câncer de pele: a segmentação das lesões de pele. Essa tarefa por si só é bastante desafiadora na perspectiva de visão computacional. Conjuntos de dados públicos não são tão extensos como para outros domínios de imagem e as anotações das imagens não são ótimas. Esses problemas têm um impacto real na performance do modelo e na sua capacidade de generalização. Ao longo desse trabalho, nós desejamos atacar a segunda questão, a qualidade das anotações das imagens. Nós analisamos as estatísticas de concordância entre anotadores no conjunto de dados de lesões de pele público mais famoso disponível e desenvolvemos algumas conclusões sobre as anotações disponíveis. Então, nós propusemos uma série de condicionamentos a serem aplicados nos dados de treino para avaliar como eles melhoram a concordância entre diferentes especialistas. Finalmente, nós analisamos como os condicionamentos afetam o treino e a avaliação de redes neurais profundas para a tarefa de segmentação de lesões de pele. Nossas conclusões sugerem que a baixa concordância entre anotadores presente no conjunto de dados ISIC Archive tem um impacto expressivo na performance dos modelos treinados, e considerar essa discordância pode, de fato, melhorar as capacidades de generalização das redes.

List of Figures

1.1 Melanomas vs Benign lesions [40] 15

1.2 Segmenting skin lesion images is challenging: samples 16

2.1 Fully Convolutional Network for Semantic Segmentation reproduced from [58] . . 24

2.2 U-Net architecture 25

2.3 Atrous convolutions explanation 26

2.4 DeepLab v3 model 27

2.5 Timeline of the development of deep learning for skin lesion segmentation 30

2.6 Basic autoencoder architecture 31

2.7 Data augmentation examples 33

2.8 Generative adversarial network training procedure 35

2.9 Overview of the segmentation approach using GAN 36

3.1 Samples of masks generated by different methods 40

3.2 Application of conditionings to samples extracted 43

3.3 Samples from the Cohen's Kappa score distribution 45

3.4 Inter-annotator agreement distributions for ISIC Archive 48

3.5 Inter-annotator agreement distributions for ISIC 2017 49

3.6 Inter-annotator agreement distributions for ISIC 2018 50

4.1 Interaction plots with test fixed with no conditioning. 57

4.2 Interaction plots with ISIC Clean as training set 58

List of Tables

2.1 ISIC 2017 Leaderboard 32

2.2 ISIC 2018 Leaderboard 32

3.1 Number of available annotations per image for each dataset. 42

3.2 Percentiles of the Cohen's Kappa mean score distributions for each dataset. . . . 46

4.1 Training and validation distributions for ISIC Full and ISIC Clean. 52

4.2 Results for models when submitted to ISIC 2018 Validation 54

4.3 Results for DeepLab V3+ when submitted to ISIC 2018 Test 55

4.4 ISIC Full vs ISIC Clean on ISIC Titans dataset 61

4.5 ISIC Full vs ISIC Clean on PH2 dataset 61

4.6 ISIC Full vs ISIC Clean on Dermofit dataset 61

A.1 Complete results of inter-annotator agreement experiments. 77

B.1 ANOVA for the experiments with inter-annotator agreement 79

Contents

1 Introduction **14**

1 How to read this text 17

2 State of the Art Review **19**

1 The Inter-Annotator Agreement Challenges for Semantic Segmentation 20

2 Deep Learning Techniques for Semantic Segmentation and Their Application to Medical Image Analysis 23

3 The Pre-Deep Learning Skin Lesion Segmentation History 27

4 Deep Learning Techniques Applied to Skin Lesion Segmentation 28

5 Generative Models for Semantic Segmentation and Their Application to Skin Lesion Analysis 34

3 Handling Inter-Annotator Agreement for Skin Lesion Segmentation **38**

1 Problem Statement 38

2 Materials and Methods 41

2.1 Dataset 41

2.2 Methods 42

3 Inter-Annotator Agreement Experimental Design 43

4 Results 44

5 Discussion 46

4 Skin Lesion Segmentation Under the Light of Inter-Annotator Agreement **51**

1 Dataset 52

2 Experimental Design 53

3 Results and Discussion 56

5 Conclusions **62**

1 Contributions 63

2 Achievements 63

3 Future Work 64

Appendices **74**

A Complete Results of Inter-Annotator Agreement Experiments **75**

B Complete Results of Inter-Annotator Agreement Experiments **78**

Chapter 1

Introduction

Image segmentation is the task of delimiting objects of interest in an image, thus separating them from other objects and from the background. In this thesis, we will focus on the segmentation of skin lesion images — which plays an important role in the automation of skin lesion analysis — where we separate the area corresponding to the lesion from the surrounding unaffected skin.

If we consider that the skin is the largest — and the most exposed — organ in the human body, we should not be surprised by the fact that skin cancer is, by large, the most common form of cancer, surpassing even prostate cancer in men, and breast and cervix cancer in women. In Brazil, it corresponds to approximately 30% of cases, or 207 770 new cases in 2018 [33].

Skin cancer is as complex and multifaceted as the skin itself. The most aggressive form, melanoma — an uncontrolled growth of the melanocytes, cells that produce skin's pigmentation — is relatively rare, but responds for a large portion of the fatalities, due to its malignancy, i.e., its tendency to spread to different regions in the body (*metastasize*, in the medical jargon). Early diagnosis is critical for a good prognosis: localized melanoma responds very well to treatment, but becomes very difficult, often impossible, to cure after it spreads. However, diagnosing melanoma, especially in its early stages, when lesions are still small, is notably difficult, even for medical specialists, as malignant and benign lesions confound easily with each other 1.1.

Another challenge for the early diagnosis of melanoma is the availability of medical specialists, as the disease incidence grows much faster than our ability to deploy newly formed dermatologists, especially in isolated, rural, or impoverished communities, where the full-time presence of those professionals might not be feasible. In that scenario, the automated detection of melanoma appears as an enticing alternative for improving the quality of care of the patients.

Due to the lack of doctors in impoverished regions in Brazil, enhancing the ability of pri-

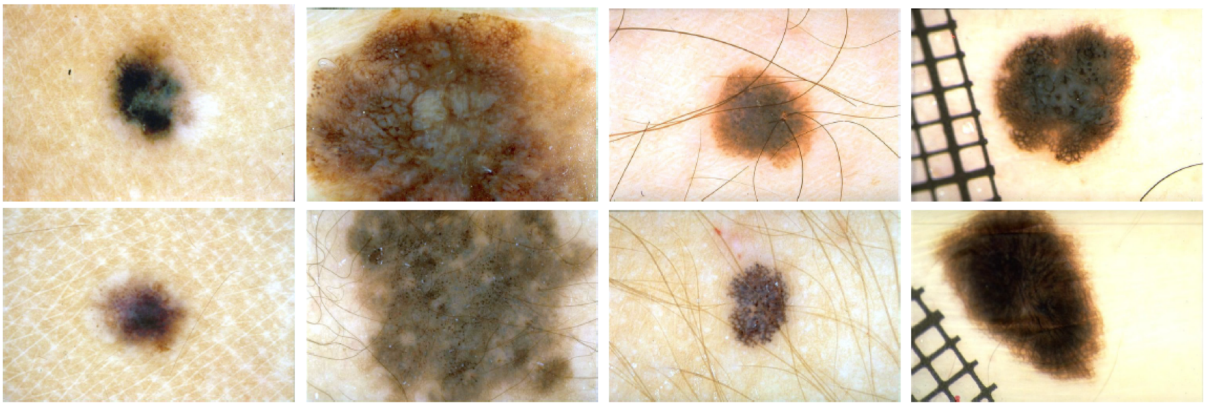

Figure 1.1: Melanomas (top row) and benign lesions (bottom row) confound with each other, since the categories present a lot of intra-class diversity, and inter-class similarity. That makes diagnosis very challenging, even for medical specilists. Reproduced from Fornaciali et al.Fornaciali et al. [40].

mary care professionals (e.g., nurses, and generalist doctors) would be a powerful tool for early diagnosis. The subject of this work — the segmentation of the skin lesion, which consists in detecting the borders/region of the lesion within a dermoscopic image — is an essential task in the pipeline of an automated screening tool.

In classical computer vision models used in the 1990s and the early 2000s, image segmentation was considered mostly an ancillary task, a necessary preprocessing step, for image classification. Such models are now obsolete for more than a decade, classification being now performed directly, without preliminary segmentation. Interest in segmentation persisted, however, with a new understanding that instead of a simple preprocessing, it was a complex semantic task *more difficult* than classification.

In automated skin lesion analysis, for example, segmentation is an invaluable tool. The complete workflow often implies locating each lesion, and even tracking lesions across images taken at different times, to measure their evolution. Those tasks strongly depend on our ability to segment the lesion. With the advent of advanced diagnostic options like full-body skin scanning, detecting and segmenting the lesions have become crucial tasks *per se*.

As mentioned, semantic segmentation is a complex task in Computer Vision, even more than classification, a challenge compounded in medical imaging due to the scarcity of training images. In addition, the annotation may be noisier, since, in many medical tasks, the boundaries of the objects of interest are fuzzy, the images may suffer from low contrast or other quality issues, and undesirable artifacts may be present. That is certainly the case for skin lesions, which often have poorly defined boundaries, and whose images present many types of artifacts: hairs, air

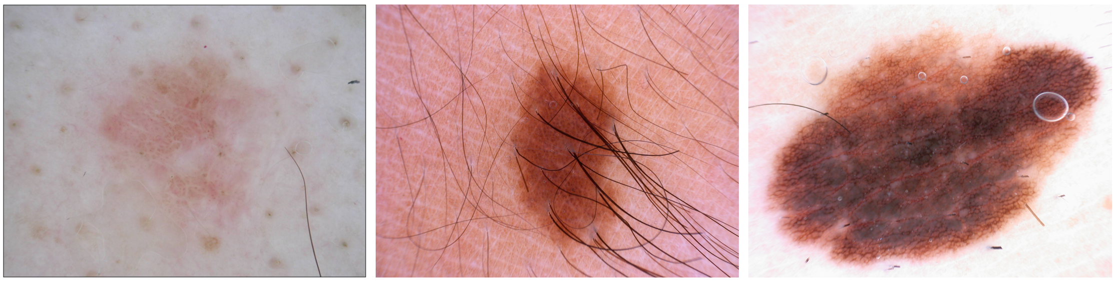

Figure 1.2: Segmenting skin lesion images is challenging due to the often fuzzy boundary of the lesion (left), and the presence of undesirable artifacts like hairs (center), or air bubbles on the fluid used on contact dermoscopy (right).

bubbles on the fluid used on contact dermoscopy, rulers and other markers, etc. (Figure 1.2).

Indeed, issues brought by annotation quality will be one of the main themes of this thesis. One important measure of annotation quality is *inter-annotator agreement*, i.e., the degree in which two independent human annotators agree on the ground-truth for the image segmentation. Our study is the very first to evaluate inter-annotator agreement for skin lesion segmentation, which we found to be troublesome low. As explored the ISIC dataset, we noticed that the annotations often diverge sharply (Figure 3.3). Characterizing precisely that divergence became the first main contribution of this thesis.

Attempting to alleviate those divergences became a second import contribution. We propose simple filters, which we call *conditionings*, that simplify the ground-truth masks, discarding noise while keeping useful information. The conditionings considerably improve the agreement between different annotations for the same lesion.

The third contribution of this work is evaluating the impact of the proposed conditionings on the task of segmentation. We will show that *discarding noise* has as considerable *positive impact* on the generalization ability of the models.

The evolution in our understanding of the role of segmentation in computer vision followed the evolution of the field. Classical models, which tended to see segmentation as ancillary preprocessing to classification, were based on the explicit extraction of features engineered by hand, such as color, shape, and texture futures. Those features would be forwarded to a separate classifier to decide on the image.

The current understanding, which sees segmentation as a semantic task, which is at least as complex as — and indeed, often *more complex* than segmentation emerged as successful models for image classification without any need for segmentation emerged, with the *bags-of-visual-words* models of the early 2000s. That perspective consolidated with the success of *deep learning* in the 2010s, as we developed end-to-end models which were able to provide very accurate image classification “directly from the pixels”, i.e., models which integrated seamlessly feature extraction and final classification.

Deep learning also allowed a sharp improvement in the performance of image segmentation. The current wave of deep generative models — networks which learn to ”generate” the distribution of the data — is an exciting frontier, not only for the creation of high-quality synthetic samples, but also for tasks like classification and segmentation. Those advances are allowing some models to outperform humans.

In this thesis, we join the state of the art on deep learning for segmentation, showing how improvements in annotation quality can have a major impact on their performance.

1 How to read this text

We organize the thesis as follows:

- In **Chapter 2**, we discuss the State of the Art of skin lesion segmentation. For the sake of completeness, we briefly present the pre-deep learning era, including some of the techniques used during the 1990s and early 2000s. The focus, however, is on current art, with the most promising and advanced segmentation methods in the medical area, including a brief overview of Generative Adversarial Networks used in our study field. The core of our SotA is a survey of works which address inter-annotator agreement, and a review of deep-learning-based skin lesion segmentation (based upon leading techniques on the latest ISIC Challenges).

- In **Chapter 3**, we analyze in-depth the inter-annotator agreement of the images of the ISIC Archive dataset. As far as we know, this is the first evaluation of inter-annotator agreement for skin lesion segmentation available. This chapter is based in Ribeiro et al. [76].

- In **Chapter 4**, we present experiments that evaluate the impact of the conditionings proposed in the previous chapter in deep-learning models for segmentation. We will show that the use of those conditionings has surprisingly positive effects on generalization.

- Finally, in **Chapter 5** we summarize our conclusions, suggesting possibilities for future works. We also list in this chapter the achievements obtained by the author of this text during his master studies.

Along with our work, we display several samples of lesion images. Unless where explicitly noted, those samples were extracted from the ISIC Archive dataset [5], which we describe in Section 1 of Chapter 3.

Chapter 2

State of the Art Review

This chapter is organized as follows: Section 1 explores the inter-annotator agreement for semantic segmentation, which is of particular relevance for us since we raise this discussion for skin lesion analysis. Section 4 surveys the central topic of this work: skin lesion segmentation using deep learning. We survey the most promising methods from the ISIC Challenge in 2017 and 2018 leaderboards. In Section 5, we advance that discussion with generative models and their application to skin lesion segmentation. To introduce this "core" material, we briefly survey deep learning for segmentation of medical images in general (Section 2), and, for the sake of completeness, we also quickly study the segmentation of skin lesion before the adoption of deep learning (Section 3).

Our research focused on improving automatic methods for skin lesion triage is vast, and the most promising ones require a significant amount of data. Like humans, computers learn by seeing real-life examples of the target subject. However, unlike us, machines still cannot generalize from small datasets. For that reason, the training of machine learning models requires a large and diverse training dataset.

Gathering medical images is challenging due to legal, economic, and technical issues. Governments are reinforcing laws on data protection [6, 2]. Industries, on the other hand, are not willing to share their private data with other players in the market to avoid losing competitive advantage [49]. Often, quality issues — either on the images or their annotations — plague existing public datasets.

1 The Inter-Annotator Agreement Challenges for Semantic Segmentation

The inter-annotator agreement is a measure of how well two or more annotators agree when attaching labels to objects belonging to specific categories. From the inter-annotator agreement analysis, we can derive different understandings about the nature of the problem. It expresses the level of difficulty of the annotation process. If two or more specialized annotators struggle to agree, we have a proxy of how hard the task is.

The inter-annotator agreement is also an expression of the reliability of the dataset under study. Data quality is a key factor when working with machine learning. When models learn from inaccurate labels, they output inaccurate predictions. Submitting the labeling task to different annotators helps to evaluate how trustworthy are the datasets and the decisions driven from the models trained with them, and gives consistency to the performance metrics achieved by the model. Martin et al. [62] discusses the major challenge when we talk about image segmentation. The question *"What is the correct segmentation?"* is much harder than *"Is this digit 5?"*, for classification. There is no unique segmentation mask for an image. Two annotators may differ in their opinions either because they perceive the content of the image differently, or because they have distinct levels of granularity, and we may not account these inequalities as inconsistencies. Segmentation evaluation can be exhausting, and the performance metrics must be aware of variations in the way annotators understand the problem and their intrinsic levels of granularity.

A third notable expression of the disagreement between annotators is related to the ambiguous character of the task. Gurari et al. [43] argues that inconsistent annotations are not only a consequence of challenging tasks and imperfect human annotators but also a consequence of inherent *ambiguity*. The original work discusses ambiguity for foreground object segmentation, especially when we have only one available ground truth for the ambiguous image. The researchers give images to a crowd and ask the people to answer the question *"If asked multiple people to draw the boundaries of a single object in the given image, do you think all people would pick the same object?"*. Each image receives five votes, and the final label comes from a majority vote. Finally, the group investigates how foreground object ambiguity impacts the way we evaluate segmentation algorithms. Samples presented along with the article shows sound outputs the algorithms produced for ambiguous images, but that disagree with the target labels. This study raises the discussion of having a single ground truth segmentation mask for an image that has a high level of disagreement between human observers.

Aroyo and Welty [14] discusses a similar issue with ambiguity in Natural Language Processing. In the original work, the research group argues that for medical relation extraction, the disagreement found in the annotation is a source of rich information. We should not see it as noise but as a signal. By addressing the idea that different annotations bring different perspectives of a crowd about the data, they state that we should not be talking about ground truth since in many cases we do not have a single correct label. We should address the crowd truth, in which disagreement is used to understand the annotated instances for training and evaluation. Instead of trying to diminish the disagreement, the scientists exploit the maximum of them based on the hypothesis that controversy exposes the vagueness and ambiguity contained in the relationships between elements of the sentences. Taking advantage of the disagreement in machine learning is not unusual. Zhou and Li [96] surveys the art of disagreement-based semi-supervised learning. Because we may have many unlabeled examples, but labeled data is scarce due to the need for human effort and expertise, semi-supervised learning tries to exploit the unlabeled data to improve performance. As the author initially explains, the key of disagreement-based semi-supervised learning is to generate multiple learners, let them collaborate to exploit unlabeled samples and maintain a substantial disagreement between the base learners.

Crowd-sourcing annotations literature extensively applies the agreement between annotators. Many researchers rely on this kind of data, even in the medical field, in which the inner complexity of the tasks are tremendous, and the annotation cost is equally extensive. Because crowd-sourced data lacks reliability, the need for measuring this gap is evident. On this topic, research using the inter-annotator agreement measure is common. Leifman et al. [54] demonstrates how to apply an approach for annotation and validation of large-scale datasets of retinal images. In the authors' words, the procedure is designed to cope with noisy ground-truth data and with non-consistent input from both experts and crowd-workers. For Machine Translation, Ambati et al. [12] also relies on crowd-sourcing for acquiring more data. The group applies active learning for text translation using crowd-sourced experts and non-experts to translate sentences. To compute the translation reliability, the group calculated the fuzzy similarity between translations given by the population and then used inter-annotator agreement as a reliability metric.

Different from crowd-sourcing based research, existing art on the inter-annotator agreement for semantic segmentation is very scarce. Contrarily to present works for lesion classification (Esteva et al. [38], Brinker et al. [21], Haenssle et al. [44]), we could not find any evaluation of annotator accuracy or inter-annotator agreement for skin lesion segmentation. Even for other tasks in medical images, systematic studies of the inter-annotator agreement are hard to find.

The most complete study we found was by Lampert et al. [53], who presents an in-depth study of the inter-annotator agreement for four image processing problems — segmentation of natural images, fissures in remote-sense images, landslides in satellite images, and blood vessels in retinoscopy — employing a large number of analytics tools to explore agreement on those tasks. The most relevant (and easy to interpret) result is the one that compares the performance of each annotator with the consensus annotation (obtained averaging the annotations). For the retinoscopy task, they had only two annotators, with Cohen's Kappa scores of 0.50 and 0.57 when compared to consensus.

Liedlgruber et al. [55] evaluate the segmentation of the hippocampus in Magnetic Resonance Image volumes for nine patients, by three different annotators, who used a graphic table to delineate the hippocampus voxels on each slice of the image. They report significant variations of agreements between the three pairs of annotators and across the nine patients, with an average 76% agreement using the Dice score, and 6.5 using the Symmetric Hausdorff distance.

Chaichulee et al. [25] report results for segmentation of areas of exposed skin on patients, aiming at non-contact vital signal monitoring. On a dataset comprising over 200 hours of video acquired from the recording of 15 preterm infants in intensive neonatal care, they asked three annotators to label the regions of exposed skin, in a semi-automated procedure where the annotator would annotate one frame. The system would attempt to propagate the annotation for the next frames, and the annotators would accept or revise the propagation. They report a mean agreement of 96.54% using the Jaccard index and also provide an estimation of the distribution of the agreements in the form of a histogram.

An extended abstract by Egger et al. [37] presents results for mandibular bone segmentation on high-resolution (512×512) 3D Computer Tomography scans. They asked two specialists to annotate the datasets and measured an agreement of 93.67% using the Dice score.

The results suggest that inter-annotator agreement for segmentation may vary widely, according to the nature of the image, and the details of the task.

2 Deep Learning Techniques for Semantic Segmentation and Their Application to Medical Image Analysis

Semantic segmentation is the task of generating dense label predictions of the pixels of an image. It is an active and challenging field of study. For medical imaging, semantic segmentation plays an important role. From our goal of skin lesion segmentation [11] to the brain and neuronal structure segmentation [60] and organs segmentation [52], these methods help physicians to understand exams better and improve their diagnosis. In this section, we explore techniques developed for different domains and became popular in the medical area. For further reading, we reference Hu et al. [47], which surveyed deep learning for cancer detection and diagnosis, and Meyer et al. [67], which surveyed the application of deep learning to radiotherapy image analyses.

Ciresan et al. [30] focused his work in the automatic segmentation of neuronal structures found in electron microscopy (EM) images. To the best of our knowledge, it is the first work to apply deep learning techniques for segmenting medical images. The proposed architecture is a series of convolutional, max-pooling, and fully-connected layers. Next, a sequence of fully-connected layers combines the outputs, and a softmax layer, in the end, guarantees a probabilistic interpretation to the output —a pixel belonging to a given class, i.e., membrane or non-membrane. The proposed work won the ISBI 2012 EM Segmentation Challenge after outperforming other techniques in three different metrics: random error, warping error, and pixel error.

Although groundbreaking, the model has two problems. 1) It is slow since it has to run for each patch of the image and the patches are highly redundant. 2) It has a trade-off between localization and context information, i.e., smaller patches improve the localization aspect but deteriorate context information, but more significant patches, which have more context information, deteriorate localization.

The work from Long et al. [58] was a breakthrough on segmenting general images from the PASCAL VOC 2012 dataset. It was the first to train fully-convolutional networks (FCN) end-to-end, pixel-to-pixel, to generate dense predictions. The network takes inputs of arbitrary size and produces correspondingly-sized outputs with efficient inference and learning.

The key idea of this work is to take well-established classification networks, e.g.AlexNet [51], ResNet [45], and GoogLeNet [82], and adapt them to semantic segmentation by replacing the fully-connected layers by fully-convolutional layers. This modification allows the network

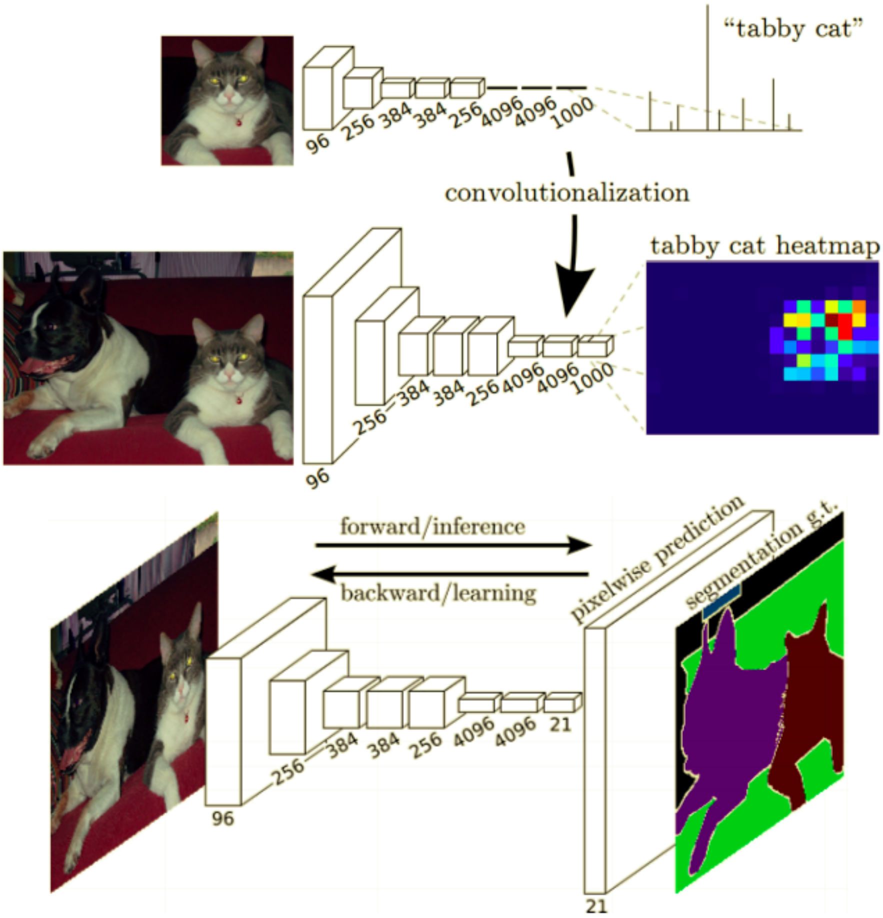

Figure 2.1: By replacing the fully connected layers of traditional classification networks by fully convolutional layers, the network can learn to make dense predictions about pixel labels. Image reproduced from [58].

to generate dense predictions in the form of a heat-map. By adding a decoding path, we transform the output back to its original size. A natural way of upsampling is by doing the inverse convolution operation, often called deconvolution, and using a spatial loss function, which enables efficient end-to-end learning. Finally, we add a skip connection to improve the localization aspects of the network.

To overcome the problems present in [30], Ronneberger et al. [77] presented the U-Net, a convolutional network for biomedical images segmentation. That architecture, based on the FCN, has a contraction path (encoder) and an expanding path (decoder). Convolutional and ReLU layers followed by max-pooling compose the encoder. The decoder is symmetrical. Concatenation-based skip connections between the encoder and the decoder help the network preserve the spatial information, a critical factor for semantic segmentation. The U-Net architecture, introduced in 2015, is still a relevant architecture for segmentation of medical images.

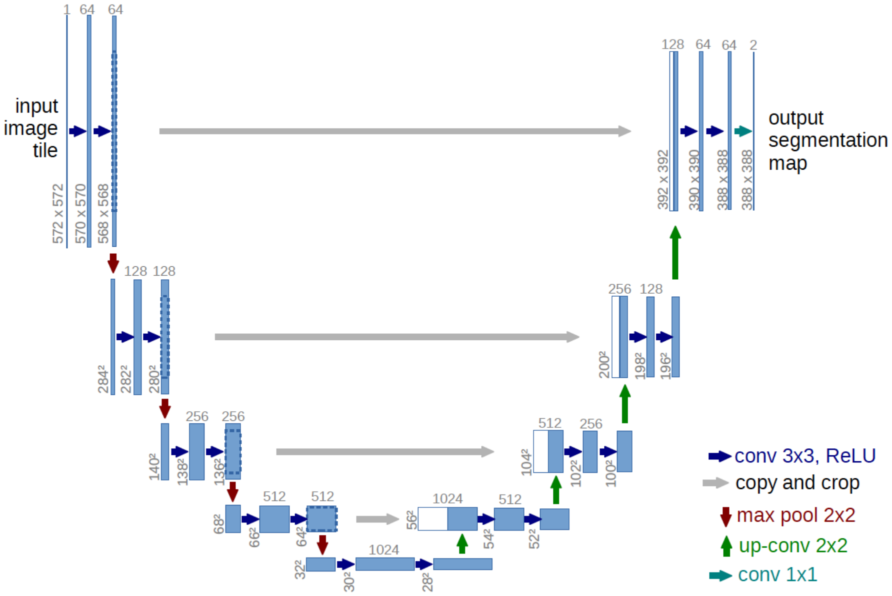

Figure 2.2: U-Net architecture (example for 32x32 pixels in the lowest resolution). The symmetry between encoder and decoder and the skip connections gives the network a U shape, from which it is named. Image reproduced from [77].

The concatenation-based skip connections enhance the problem of vanishing gradient, when the gradients of the most initial layers get so close to zero during backpropagation that the weights do not update anymore and the network stops learning. Overcoming this problem is a challenge to enable deeper architectures, a key factor for more robust and accurate algorithms.

On this topic, Quan et al. [74] proposed the FusionNet, a fully-residual convolutional network for image segmentation of connectomics. The main differences between FusionNet and the traditional U-Net are that the first replaces the original skip connections by sum-based ones and it introduces another skip connection inside the residual blocks, which gives the network a fully-residual fashion. The article demonstrates the flexibility of the architecture for two medical image segmentation tasks: cell membrane segmentation and cell nucleus segmentation.

In 2017, Chen et al. [27] proposed the DeepLab, a different approach for semantic segmentation. As originally described, the new method brings three contributions to state of the art. First, it introduces atrous convolutions (Figure 2.3) as a tool for making dense predictions. The advantage of atrous convolutions over regular ones is that by adding spacing over the convolutional kernel, we increase the receptive field, and consequently learn richer context information, without adding more complexity to the operation. Second, it uses atrous spatial pyramid pooling (ASPP) to segment objects at multiple scales robustly. Finally, the paper combines the results of deep convolutional neural networks (DCNN) with fully connected conditional random fields (CRF) [58] to improve localization of the output.

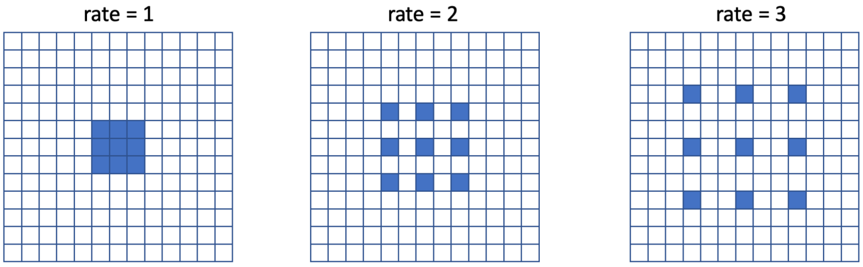

Figure 2.3: The idea behind atrous convolutions. By adding spacing inside the convolutional kernel, it is possible to increase the knowledge about context without increasing the complexity of the model. Image based on [10].

Later in the same year, Chen et al. [28] revisited the idea of atrous convolutions with a new version of the DeepLab which, in the author's words, explicitly adjusts filter's field-of-view as well as control the resolution of feature responses computed by DCNN. The new architecture improves the performance on the benchmark datasets when compared to the previous version without the need for CRF post-processing. After extensive experimentation, the most recent version of DeepLab, the DeepLab V3+, became the leading architecture for our experiments once it over-performed all the candidate models on the ISIC 2018 test dataset.

Not only for general semantic segmentation tasks, but also other medical imaging tasks, the DeepLab model has been achieving excellent results, and many different works apply it as a baseline. Chen et al. [29] used a DeepLab like architecture for a multi-task framework on skin lesion segmentation and selected the original one as the baseline for the paper. Bai et al. [15] used a DeepLab-based architecture for semi-supervised learning in cardiac MRI segmentation. Finally, as we will see later, the architecture proved itself as groundbreaking during the 2018 edition of the ISIC Challenge.

In 2019, Liu et al. [57] proposed a Neural Architecture Search (NAS) method for semantic segmentation named Auto-Deeplab. The NAS method proposes to automatically design the neural network architecture, minimizing the need for human efforts. This work is innovative when compared to previous NAS ideas once it proposes hierarchical architecture search space by searching both the network level structure and the cell level structure. The model achieves good results without using any ImageNet [34] pre-training. When compared to the original DeepLab V3+, the Auto-Deeplab performs slightly worse in the benchmark datasets. The results obtained

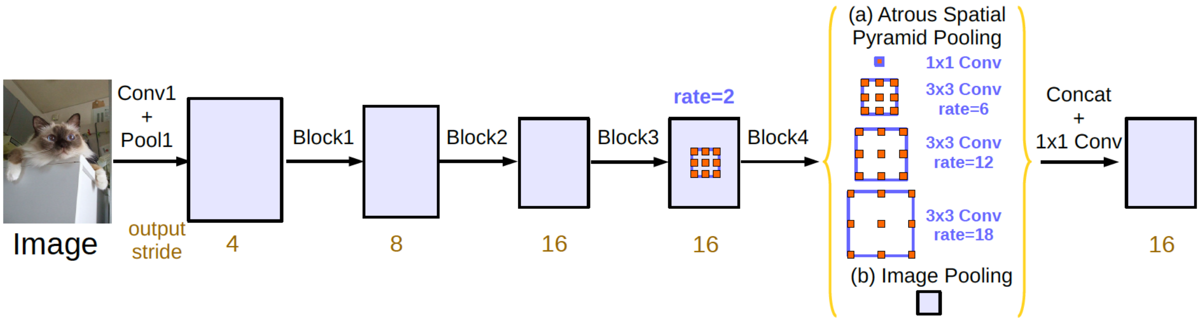

Figure 2.4: DeepLab v3 model: Parallel modules with atrous convolution (ASPP), augmented with image-level features. Reproduced from [28].

by the research team is consistent with the results we achieved when submitting both networks to skin lesion segmentation.

3 The Pre-Deep Learning Skin Lesion Segmentation History

As said before, researchers have been trying to improve skin lesion segmentation techniques for a long time, but the most promising methods came with deep learning in the 2010s. Before deep learning, color and texture information composed the basis of automated segmentation approaches. Umbaugh et al. [84, 85] applied color based algorithms to task. Green et al. [42], Dhawan and Sim [36] and Moss et al. [68] added texture extractors to the equation. Sahoo et al. [78] compared different common computer vision methods adopted in the 1990s for similar tasks. In this section, we discuss these methods. During this time, little to no work was developed using supervised techniques. Celebi et al. [23] analyzed 16 articles focused on skin lesion segmentation, and just two of them used supervised methods.

We limit our analysis in this section to an overview of the field and promising methods proposed before deep learning. For a more comprehensive view, we reference a survey published by Celebi et al. [24] that presents an overview of 50 published articles describing the state of the art of border detection algorithms. The survey reviews the pre-processing, segmentation methods, post-processing, and evaluation criteria of several works related to the area. It then presents a comparison of the methods concerning different aspects.

In 1999, Xu et al. [89] applied a three-step method for the segmentation of the skin lesions. At first, the method transforms the image from the RGB color space to the CIE L*a*b color space. The second step consists of generating an initial estimation of the lesion border and location. To do so, they pass a low-frequency filter to the image, removing noise caused by the presence of artifacts in the lesion that deteriorates the image. Then, they apply a global optimal threshold value to estimate the initial lesion border. Finally, the third step consists of refining the border estimative using the closed elastic border method according to a local optimum threshold value.

A few years later, Rajab et al. [75] described two different methods to address skin lesion segmentation. The first applies an iterative method to separate lesion from background skin. The second applies a multilayer perceptron trained with 3x3 pixels border patterns to detect lesion edges.

During the first years of the 2000s, Celebi et al. [22] employed a modification of the JSEG algorithm [35] that consists of three steps: a pre-processing step to remove image artifacts, a color quantization step and a post-processing step to remove remaining healthy skin from the generated segmentation mask. The difference between the work of Celebi and the original JSEG is that the first uses a median filter during the pre-processing instead of Peer Group Filter applied by the former in order to better remove lesion artifacts.

In subsequent years, other scientists tried different techniques to improve border detection. Yuan et al. [91] used a narrow band graph partitioning method. Naz et al. [70] describes several articles on the fuzzy clustering technique, when each data point has a probability of belonging to a given class. Schaefer et al. [79] uses color enhancement to improve the segmentation generated by the method described by Rajab et al. [75]. Moreover, Zhou et al. [95] applied a gradient vector flow with the mean shift to segmentation of skin lesions.

However, most of the work done during the pre-deep learning time was mainly on unsupervised approaches, Wighton et al. [88] described in 2011 a supervised method that goals to generalize common subtasks of skin lesion diagnosis. The proposed work aims to generalize the lesion segmentation, hair detection, and pigmented network detection tasks.

4 Deep Learning Techniques Applied to Skin Lesion Segmentation

This section is the core of our literature review, surveying the works which are the closest to ours: those who employ deep learning for skin lesion segmentation. With more computational resources available, the deep learning era came to the skin lesion analysis enabling high performance with more robust methods. The focus of this section is mainly the ISIC challenge — a competition hosted every year that challenges its participants to improve results on different tasks related to skin cancer. We work through the leaderboards of the last two editions, examining the top-ranked approaches. For further reading, we refer the reader to [11], which presents a survey of the state-of-the-art algorithms and techniques for performing skin lesion segmentation. Less recent but still relevant, Fornaciali et al. [40] survey, analyze and criticize the art of melanoma screening. Finally, Ammar et al. [13] proposed a 31 layers deep architecture which achieves high accuracy segmentation for both PH2 dataset and ISIC 2017 dataset.

Not only for skin lesion segmentation but computer vision as a whole, the central contribution of deep learning is that it does not rely on handcrafted features. The neural network is trained from the raw pixels of an image and learns to detect all kinds of patterns, bringing reliability to the methods and improving generalization.

Neural network models are a way of representing highly non-linear functions understandably and naturally. It tries to mimic the neural system of animals with the metaphor of inputs activating neurons to generate an output. Adding more neurons (and connections) to the network enable the representation of more complicated functions to fit complex data. Deep neural networks are just like shallow and traditional ones, but with a more significant number of neurons distributed in several layers.

The problem with this technique is that with a deeper architecture, the neural network has to learn a higher number of parameters. Moreover, with more parameters, training requires massive datasets. As we already discussed, data is a finite and scarce resource when it comes to medical images, which means that we need to find solutions to enable learning.

On the international collaboration towards melanoma detection, the ISIC community started a competition to challenge scientists, researchers and AI developers all over the world to develop methods to improve results on different tasks related to skin lesion analysis. Within the competition, the community built the ISIC dataset, and nowadays it is one of the essential sources of skin images. Along with this work, we deal mainly with the 2017 [31] and 2018 [4] versions of it. In this section, we will review several deep learning methods developed for skin lesion semantic segmentation and walk through the leaderboard of the last two years of the ISIC challenge.

With some particularities, almost all semantic segmentation architectures follow the autoencoder architecture seen in Figure 2.6. The traditional autoencoder consists of an encoder and symmetric decoder, with the latent space in the middle. This architecture is instrumental for several computer vision tasks, but it has a problem with semantic segmentation. It loses spatial information during the encoding path. As seen before, the most critical architectures for segmentation of biomedical images includes skip connections between analogous layers of the encoder and the decoder. These connections help the models to reconstruct spatial information and generate the probability map.

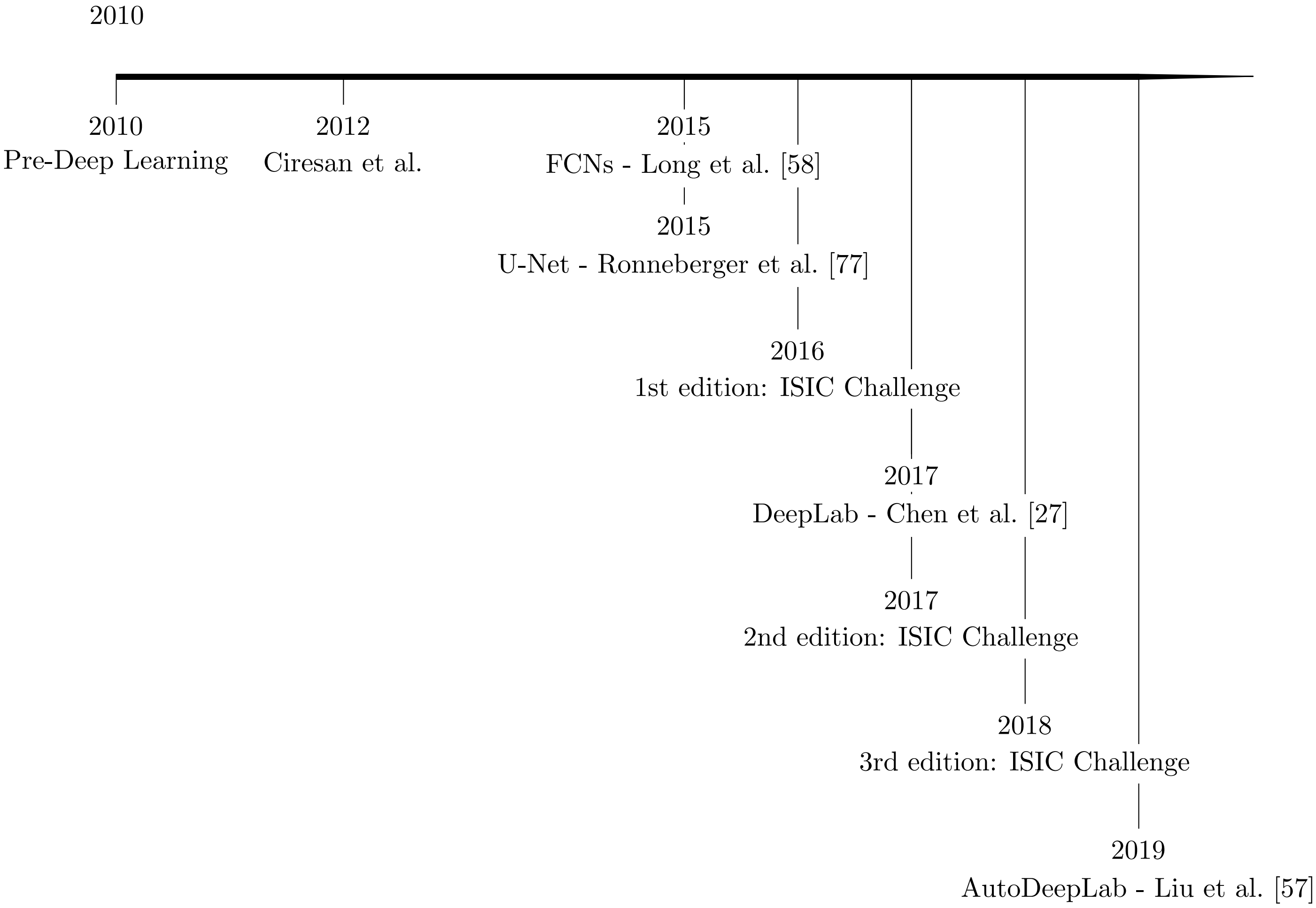
2010
2010
Pre-Deep Learning
2012
Ciresan et al.
2015
FCNs - Long et al. [58]
2015
U-Net - Ronneberger et al. [77]
2016
1st edition: ISIC Challenge
2017
DeepLab - Chen et al. [27]
2017
2nd edition: ISIC Challenge
2018
3rd edition: ISIC Challenge
2019
AutoDeepLab - Liu et al. [57]

Figure 2.5: Timeline of the development of deep learning for skin lesion segmentation.

Segmentation networks usually do not differ a lot in encoding path. Traditional encoder architectures commonly used for classification like AlexNet, ResNet and GoogLeNet are the basis for these architectures. The main difference in segmentation networks are the decoding path, that may apply different upsampling techniques, and different types of skip connections. Training both the encoding and the decoding path from scratch is an arduous task and needs lots of data. For overcoming the unavailability of data, scientists developed a technique called Transfer Learning [71], which consists of transferring the knowledge acquired during the training of a general task to the performance of a different one. This method is widely used in deep learning and helps improving performance on tasks with scarce data. Many encoder architectures are pre-trained in the ImageNet [3] dataset and fine-tuned for skin lesion segmentation. Although less common, it is possible to apply transfer learning to the full encoder-decoder path by transferring

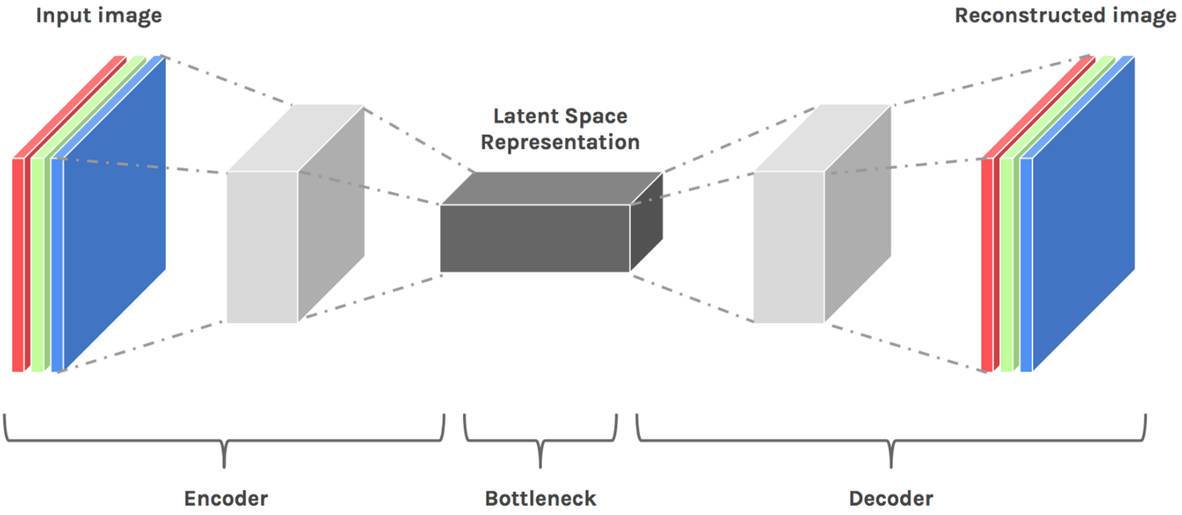

Figure 2.6: The most basic autoencoder architecture consists of an encoding path and a decoding path, with the latent space in the middle. Image reproduced from [8].

knowledge acquired from other semantic segmentation datasets, e.g.PASCAL VOC 2012 and COCO.

As discussed before, the U-shaped architecture is one of the most influential topologies for biomedical images, especially for skin lesion images. Our research group [66] achieved the 5th place and Berseth [17] achieved second place in the ISIC 2017 using an U-Net-like network. The first ensembles four models. Two trained with the 2 000 samples of the challenge training set, without a validation split, for 250 and 500 epochs respectively, and two trained and validated with a 1 600/400 split for 220 epochs. The second applied the pure U-Net model and extensively applied distortions to the challenge data, going from 2 000 images to 20 000 images. During the 2018 edition of the ISIC challenge, the U-Net was also present in the leaderboard. Koohbanani et al. [50] used a modified version of the network to achieve the 5th place in the competition.

Other architectures were also very competitive during the 2017 competition. Yuan [92] applied a fully convolutional-deconvolutional network with ReLU activation function in the convolutional and deconvolutional layers to achieve the 1st place. The group not only used the RGB channels as inputs of the network but also the three channels of the Hue-Saturation-Value space and the L channel (lightness) of the CIE L*a*b space. Bi et al. [18] applied deep residual blocks (ResNet [45]) to achieve the 3rd place. The researchers used both the challenge data and ISIC-archive data, reaching a total of 9 800 images.

During the 2018 edition of the ISIC Challenge, other architectures had a great performance. Qian et al. [73], winner of the competition, applied a two-stage method for segmenting the

ISIC 2017 Leaderboard						
Rank	Competitor	Accuracy	Dice Coefficient	Sensitivity	Specificity	Jaccard Index*
1	Yuan et al.	0.934	0.849	0.825	0.975	0.765
2	Berseth et al.	0.932	0.847	0.820	0.978	0.762
3	Bi et al.	0.934	0.844	0.802	0.985	0.760
4	Bi et al.	0.934	0.842	0.801	0.984	0.758
5	Tavares et al.	0.931	0.839	0.817	0.970	0.754

Table 2.1: ISIC 2017 Leaderboard. Jaccard index, marked with *, is the main metric for the competition. The 5th place, Tavares et al.[66], refers to our research group submission.

ISIC 2018 Leaderboard				
Rank	Competitor	Use external data	Jaccard Index	Jaccard Index (0.65 threshold)*
1	Qian et al.	No	0.838	0.802
2	Du et al.	No	0.837	0.799
3	Ji et al.	No	0.834	0.799
4	Xue et al.	No	0.837	0.798
5	Koohbanani et al.	0.836	No	0.796

Table 2.2: ISIC 2018 Leaderboard. Jaccard index with 0.65 threshold, marked with *, is the main metric for the competition.

lesions. At first, they applied a MaskRCNN [46] to detect the lesion location in the image and then applied an encoder-decoder architecture inspired by the DeepLab [27] and the PSPNet [93] architectures. The 2nd place used the DeepLab architecture with transfer learning from the PASCAL VOC 2012 dataset. As a post-processing technique, the group applied Conditional Random Fields [94] to refine the output mask. The 3rd place [92] used a traditional encoder-decoder architecture with a ResNet [45] as the encoder network and a sequence of deconvolutional layers as the decoder. Table 2.1 and Table 2.2 summarizes the leaderboards of the last two editions of the ISIC challenge.

All of the described works apply data augmentation, which consists of applying small distortions to the input image in order to generate new samples for the training set. Widespread techniques applied are rotation, flipping, zooming, and shifting, among others. Figure 2.7 shows some examples of augmented images. Data augmentation not only enriches the dataset but also makes the model more robust to perturbations.

The technique is not useful *only* for image segmentation. Perez et al. [72] evaluated the performance of three different Convolutional Neural Networks (Inception-v4 [83], ResNet [45] and

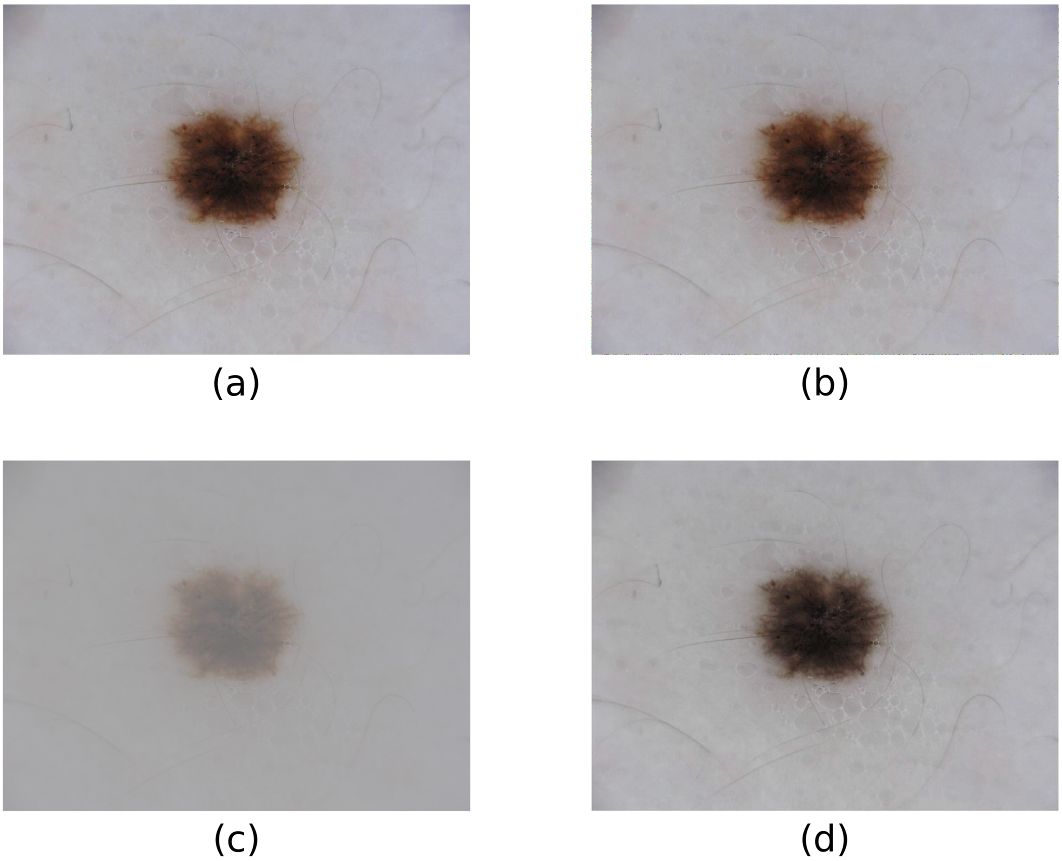

Figure 2.7: Augmented samples generated using the techniques we used in our final work. Images a-d presents the same lesion image with the following configuration: (a) original image, (b) Gaussian noise, (c) contrast degradation, and (d) color degradation.

DenseNet [48]) on lesion classification when submitted to 13 different augmentation techniques. The proposed work resulted in better performance for classification than the top 3 submitters of ISIC 2017 competition without using additional data.

Our research group also joined the ISIC 2018 Challenge for all the three tasks: lesion boundaries segmentation, lesion attributes segmentation, and lesion classification. During the challenge, the present author contributed mainly to the first of the three. Discussing the methods used for the other two is beyond the scope of this study. If interested, we reference the reader to our technical report [19] that describes all of the approaches we tested during the competition.

For the challenge, we decided to keep the U-shaped networks from the previous participation of the group in the challenge. We tested two models: a traditional U-Net-like network with a VGG-16 encoder pre-trained on the ImageNet dataset and the FusionNet [74], which has a fully-residual architecture and performs well for segmenting connectomics images.

We trained our models on two different datasets: the challenge data and the challenge data plus external data, extracted from the ISIC Archive dataset [5]. What we learned from this training configuration is that for lesion segmentation, the quality of the data and its targets is more relevant than the amount of data used during training. The segmentation masks used as ground truth in both datasets have a sizeable inter-human variability, caused by the differences in the methods for generating them. With less data, we reduce the variance between the ground truths and the algorithm generalizes better. This conclusion is consistent with other participants, which reported that adding more data degraded the algorithm's overall performance.

An essential tool that boosted our performance during the competition was the Cyclic Learning Rate technique [81]. The method consists of varying the learning rate cyclically within reasonable boundaries during training, which improves the model accuracy and reduces the training time by preventing the model to stick to local minima. For the loss function, we worked with a combination of the Binary Cross Entropy function and a variation of the Jaccard index function.

We submitted three configurations of our models: 1) average of FusionNet trained on Challenge data only, and U-Net trained on Challenge data only; 2) average of FusionNet trained on Challenge data only, U-Net trained on Challenge data only, and FusionNet trained on Challenge data and external data; 3) U-Net trained on Challenge data only. Our official results on the official test set were, respectively, 0.694, 0.686, and 0.728 for the threshold Jaccard index. Also, our positions of each submission were, respectively, 88th, 93th, and 56th among 112 submissions.

5 Generative Models for Semantic Segmentation and Their Application to Skin Lesion Analysis

Goodfellow et al. [41] first introduced the Generative Adversarial Networks in 2014. The proposed framework goals to solve common difficulties related to deep generative models. As argued in the original work, these models have had less impact due to the difficulty to overcome the problem of approximating many intractable probabilistic computations that arise during maximum a posteriori estimation and due to the difficulty of leveraging the benefits of piecewise linear units in the generative context.

On the adversarial framework, two models are supposed to compete with each other: the generative network (G) and the discriminative network (D). While the task of G is to generate the most realistic samples as possible, the task of D is to tell whether the input sample came from the data distribution or the generative model distribution.

In the cited work, Ian Goodfellow proposes a simple metaphor for the adversarial framework. We can see the generative network can as a team of counterfeiters, which are willing to produce fake currency and use it as real ones. In this scenario, the discriminative network works as the police, trying to separate the real money from the fake. The adversarial framework is a two-player game in which both teams try to improve their methods until the fake samples and the original ones are indistinguishable from each other, and the probability of D predicting that a given sample came from the data distribution is equal to 0.5.

Figure 2.8 presents a straightforward explanation of the training procedure in the adversarial framework. GANs are trained by simultaneously updating the discriminative network so that it learns to discriminate samples of the data distribution (p_x) from samples of the generative one (p_g). First, p_x and p_g are close to each other, but they are not the same. We also have a poor classifier to predict whether a given sample came from the first or the second. In the inner loop of the algorithm, D learns to discriminate the samples, converging to the optimal discriminative distribution. We then train the generative model to draw samples closer to the data distribution, fooling the discriminator. After enough iterations, the generative model cannot improve anymore once p_x and p_g are indistinguishable from each other, i.e., $D(x) = 0.5$.

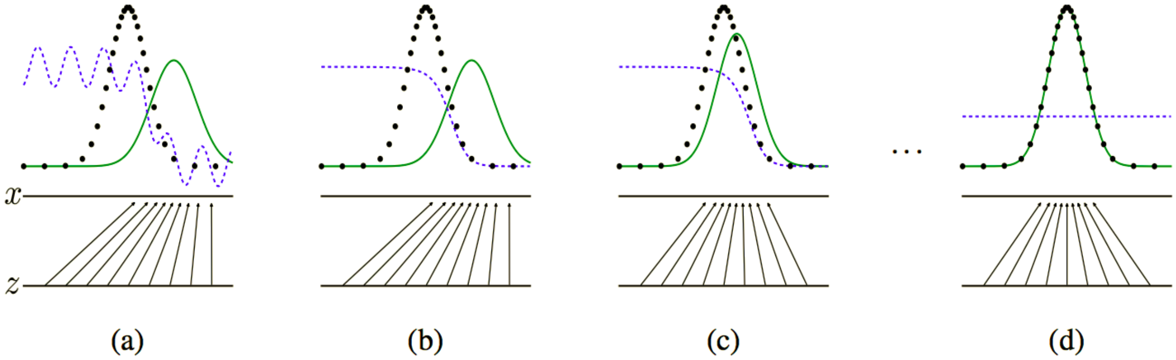

Figure 2.8: The image presents an explanation of the generative adversarial network training procedure. The black dotted line represents the data distribution (p_x), the solid green line represents the generative distribution (p_g), and the blue dashed line represents the discriminative distribution. Image reproduced from [41]

Based on the framework proposed by Goodfellow, Luc et al. [59] was the first to explore the adversarial training approach for semantic segmentation, to the best of our knowledge. The approach has two advantages when comparing to previous methods. First, it proves that adversarial training is flexible enough and has a high capacity of detecting an extensive range of

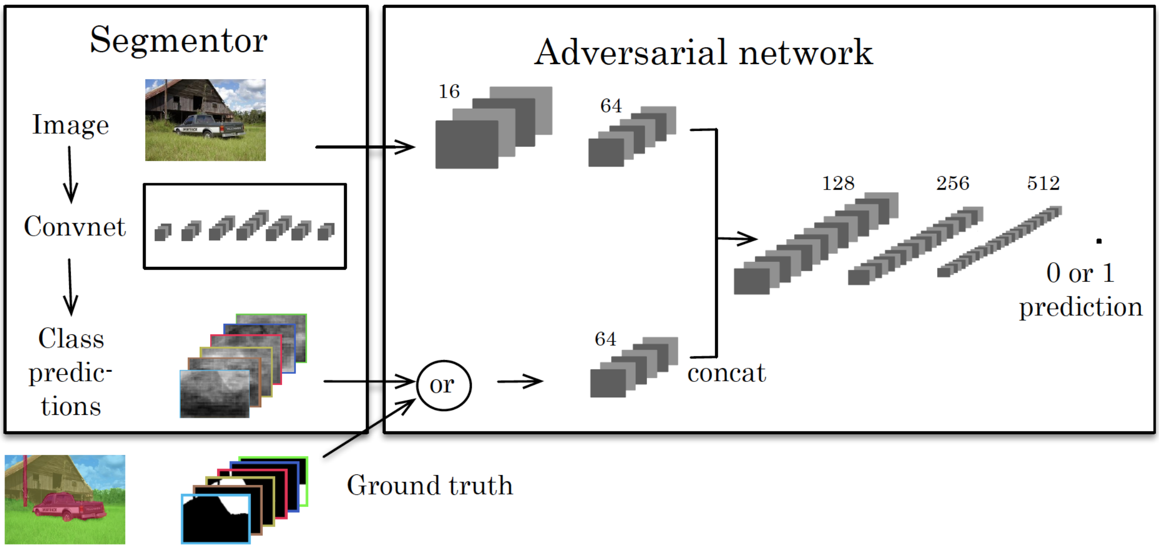

Figure 2.9: An overview of the segmentation approach proposed by Luc et al. [59]. Left: The generative network receives an RGB image and produces the segmentation masks. Right: The discriminative network receives per pixel label maps and produces a class label (0 for synthetic mask and 1 for ground truth). The discriminative networks optionally receive the RGB image as well.

probability distributions available in the data. Second, once trained, the model is very efficient since it does not rely on higher-order terms.

For semantic segmentation, the generative framework has a subtle difference from the presented above. The task for the discriminator network is not to predict the probability of the input image belonging to the dataset. Instead, its task is to distinguish between the output image and the ground truth. With the metaphor used before, the task is to predict which currency is real and which one is fake given two coins, instead of predicting if a given coin is real or fake.

For training the network, the group optimizes an objective function that combines a conventional multi-class cross-entropy loss with an adversarial term. The adversarial term encourages the model to produce segmentation maps that cannot be distinguished from the ground truth by an adversarial binary classification model [59].

When we talk about skin lesions, generative methods can have a significant impact. On our main task, Xue et al. [90] proposed the SegAN, end-to-end adversarial network architecture with a multi-scale loss for segmenting biomedical images. The adversarial training proposed not only improved state of the art on the ISIC 2017 dataset but also did not suffer from unstable training as other adversarial networks. Using the SegAN approach, the group participated in the ISIC 2018 skin lesion analysis challenge and got the 4th position among 112 submissions, which shows that the adversarial training is not only an exciting approach for skin lesion segmentation, but also it is a very competitive one.

On the other hand, generative models are a way of overcoming the dataset size issue. Bissoto et al. [20] employed the pix2pixHD GAN [86] to combine the semantic map, which corresponds to the segmentation mask used on previously discussed works, and the instance map, an image where each pixel combine information from its class and its instance, of different images to generate high-resolution images of skin lesions never seen before. This work shows up as an up-and-coming technique for enriching skin lesions datasets without the need for human data extraction and manual annotation.

Chapter 3

Handling Inter-Annotator Agreement for Skin Lesion Segmentation

We base this Chapter in our recent work [76]. We explore the issue of the inter-annotator agreement for training and evaluating automated segmentation of skin lesions. We explore what different degrees of agreement represent and how they affect different use cases for segmentation. We also evaluate how conditioning the ground truths using different (but elementary) algorithms may help to enhance agreement and may be appropriate for some use cases.

We conducted our experiments on the ISIC Archive — the most massive public dataset of skin lesion images accompanied of reference segmentation by humans — and as far as we know, the only one to provide more than one reference segmentation per image. The ISIC Archive is the baseline for most of the research in the area [47, 90].

1 Problem Statement

The segmentation of skin lesions is a cornerstone task for automated skin lesion analysis, useful both as an end-result to locate/detect the lesions and as an ancillary task for lesion classification. Lesion segmentation, however, is a very challenging task, due not only to the challenge of image segmentation itself but also to the difficulty in obtaining properly annotated data. Detecting the borders of lesions with high accuracy is challenging even for trained humans, since, for many lesions, those borders are fuzzy and ill-defined.

Since the inception of automated skin lesion analysis, the segmentation of lesions has attracted scientific interest [36, 68]. Early methods of lesion classification tended to strictly mimic medical criteria [40], such as the ABCD rule [69], in which both (B)order irregularity and large (D)iameter depend on lesion segmentation to be estimated automatically. Such methods were also consonant with the art on computer vision of the 1990s, in which segmentation was often considered a crucial preliminary step for classification (e.g., to allow extracting shape features).

The transition of computer vision art to bags-of-words models in the 2000s [80] and deep learning in the 2010s [51] spelled the end of the viewpoint of segmentation as an ancillary technique in preparation for classification. That understanding, however, also increased the appreciation of segmentation for its own merits. With the accumulated experience brought by collective efforts like the PASCAL VOC [39] and the ImageNet [34] challenges, we now understand not only that one can tackle segmentation and classification independently, but also that segmentation is usually *much more challenging* than classification.

Those advances in computer vision appear in the contemporary art in skin lesion analysis [40, 65, 19, 72], in which, although lesion segmentation is sometimes still used to help in the classification, the community understands it as an essential and challenging task in itself.

Obtaining accurate annotations is paramount for all machine learning techniques. The accuracy of annotations imposes an upper bound on the *actual, real world* accuracy of learned models. Although, in theory, any model can reach 100% of accuracy on any dataset, accuracies above those of the annotations only reflect the ability of models of learning the datasets' biases. Thus, appraising annotation accuracy is vital to decide the point above which it becomes counterproductive to keep working on the models. Estimating annotation accuracy is often, however, impossible, since it requires, in principle a *more reliable* standard than the one provided by the ground truths themselves. In scenarios where such a standard is not available, the inter-annotator agreement can act as a proxy estimation.

Because of the complicated procedure of annotating borders and regions (in comparison to just providing a label) and the often subjective nature of the task, in which the position of a border/limits of a region may be ill-defined (Figure 3.1), segmentation, especially, brings challenges for annotation accuracy.

A vital consideration to appraise the impact of annotation accuracy for segmentation is its intended use. For skin lesions, we can quickly identify at least three very distinct use cases, with progressively stricter demands of accuracy:

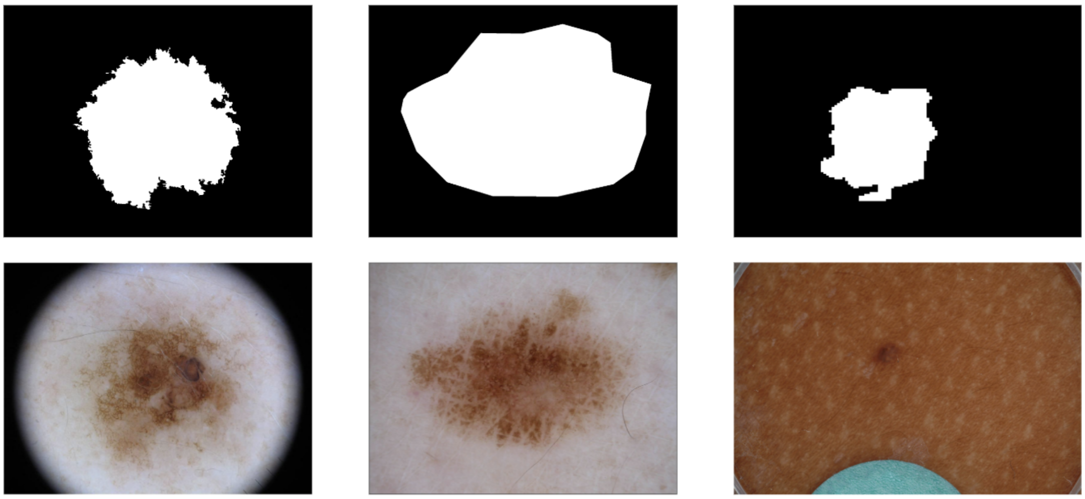

Figure 3.1: Top: flood-fill algorithm controlled by the annotator. Middle: manual polygon tracing. Bottom: fully-automated annotation validated by a human annotator.

- **Localization:** here we are interested in *detecting* the presence of the lesions, and *locating* its position. The precise limits of the lesion are not important. For this use case, an approximate bounding box may suffice, or even less: a single point anywhere inside the lesion may be enough. This level of annotation may be useful, for example, for automatically locating the lesions in a full-body skin exam.

- **Demarcation:** here we must not only locate the lesions but also correctly determine their *overall shape*. We want to be able to estimate metrics such as the lesion diameter, eccentricity, and overall symmetry.

- **Description:** here we want to fully characterize the lesion border, including detailed characteristics such as smoothness vs. irregularity. This level of annotation is the one required to mimic the medical algorithms (e.g., the ABCD rule) straightforwardly.

The list above does not intend to be exhaustive; it means to illustrate how different use cases may impose very different demands to both the ground-truth annotators and the automated techniques.

In this work, we will discuss the impact of different levels of inter-annotator agreement on those use cases, and explore how very simple conditionings may significantly improve the agreement for some use cases. Our main contributions are:

- An estimation of the inter-annotator agreement for skin-lesion segmentation. We not only provide simple statistics (such as a mean) but instead attempt to characterize the distribution of the agreements fully;

- A visual presentation of representative samples for different agreements, in order to help the reader to grasp their qualitative meaning;
- An evaluation of several simple procedures that may help to improve inter-annotator agreement if used to condition the ground truths. Those conditionings may be helpful for some use cases.

2 Materials and Methods

2.1 Dataset

The experiments in this chapter are based in the ISIC Archive [5] — curated by the International Skin Imaging Collaboration — the largest publicly available dataset of images of skin lesions, with over 23 000 annotated images. Although a few other datasets also provide segmentation information [16, 64], as far as we know, the ISIC Archive is the only public dataset with more than one segmentation annotation per lesion, and thus the only one where we can appraise inter-annotator agreement.

At the time we ran our experiments, the ISIC Archive dataset contained exactly 23 907 images of lesions, 13 779 of which had segmentation ground truths. For our study, however, we need images with at least *two* ground truths, reducing those to the much smaller subset of 2 233.

The ISIC Challenge employs a subset of the ISIC Archive, which included a task for lesion segmentation [61, 31, 32]. Since the challenge allowed for the first time the researchers to directly compare their techniques in a fully reproducible setting, it has been very influential in the community. Therefore, in addition to analyzing the full archive, we also explored the image subsets used in the past two challenges to see if there were any appreciable differences. Table 3.1 summarizes all three datasets.

The ground truth annotations in the ISIC Archive are highly variable. Just considering the subsets used for the challenges, there are already three different methods to create the annotations. As stated by the challenge organizers [4]: (1) a semi-automated flood-fill algorithm, with parameters chosen by a human expert; (2) a manual polygon tracing by a human expert; (3) a fully-automated algorithm, reviewed and accepted by a human expert. As shown in Figure 3.1, the first method tends to create a very irregular border, the second very smooth borders, and the third is in-between, with borders that appear “pixelated”.

Annotations	ISIC Archive	ISIC 2017	ISIC 2018
1	11 546	1 290	1 488
2	2 094	616	995
3	100	67	71
4+	39	27	34
Total	13 779	2 000	2 588

Table 3.1: Number of available annotations per image for each dataset.

2.2 Methods

In this work, we not only measure the inter-annotator agreement on the ground truths but also evaluate how simple conditioning of the ground truths may help to enhance that agreement.

The conditioning consists of applying simple image processing operations to all ground truth masks. The proposed conditionings are very straightforward and deterministic — there is no learning involved. We list them below:

- **Opening:** this is a morphological operation that removes details from the *foreground* (lesion). The structuring element was a square of five pixels;
- **Closing:** this is a morphological operation that removes details from the *background*, e.g., small holes or tears. Same structuring element as above;
- **Convex hull:** here we find the smallest convex shape that covers the entire lesion;
- **Opening or Closing + Convex hull:** the morphological operation followed by the convex hull;
- **Bounding box:** here we find the smallest rectangle with sides parallel to the image that covers the entire lesion.

Figure 3.2 illustrates those operations. From a theoretical point of view, one may interpret the conditioning as **denoising** operations, whose aim is to preserve the cogent information about the lesion segmentation, while discarding details which depend on the choice of one particular annotator.

We implemented all of the conditionings in Python 3. Apart from the bounding box, which was developed from scratch by our team, we extracted all of the conditionings and structuring

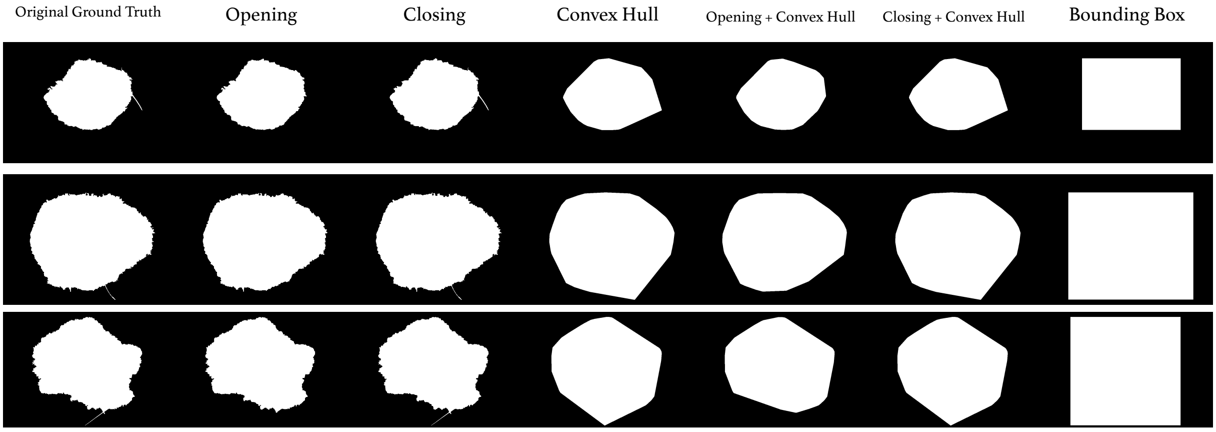

Figure 3.2: For each sample, we present its corresponding mask conditioned with opening, closing, convex hull, opening + convex hull, closing + convex hull, and bounding box. Note how the opening can remove small details from the foreground, which may significantly affect the convex hull.

elements from the morphology package of the scikit-image library [9]. Auxiliary code was developed using the numpy library [7]. The code we used to both condition the ground truths and to analyze the results is available at our Github repository[1].

3 Inter-Annotator Agreement Experimental Design

There are many metrics available to evaluate the level of agreement between two annotations, e.g. Jaccard Index, Dice Coefficient, and Cohen's Kappa Score. In our experiments, we employ the third [63], which offers, over the alternatives, the advantage of taking into account the probability of the agreement occurring by chance. Equation 3.1 presents its the mathematical formulation.

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}; -1 \leq \kappa \leq 1 \tag{3.1}$$

In the referred equation, p_o refers to the relative observed agreement between raters, p_e refers to the hypothetical probability of chance agreement. The score ranges from -1 to 1, is zero for pure chance, positive for better than chance, and negative for worse than chance.

Figure 3.3 displays examples of what different levels of Cohen's Kappa Score mean. For each original image, we have two annotations provided by different annotators, that we show

[1] https://github.com/vribeiro1/skin-lesion-segmentation-agreement

immediately on the right of the original image. The two annotations are superimposed (one in blue, the other in yellow), and we add some transparency so the two can appear. The area with mixed colors represents the overlap between the two annotations. We add in the bottom right of each image the Cohen's Kappa Score between the two masks. The images are sorted top-down in ascending order of Kappa.

The images in the top, with lower levels of agreement, have very different annotations. The first two rows (four images), which have a negative score, have no intersection. The agreement is worse than chance.

The following images have a positive score. The third row ($0.0 < \kappa < 0.5$) have a minimal intersection; there is a large area of disagreement. For the images with $\kappa \approx 0.5$, we see that the two annotators disagree by the level of granularity, or the method applied. For the last three rows, with high agreement, the annotations are equivalent, and the disagreement is minimal.

For a given lesion, we compute the Cohen's Kappa Score between its ground truth annotations. If a lesion has more than two ground truths, we take the average of the Kappa of all possible pairs. We tabulate all Kappas to estimate the distribution of the values (and associated statistics) for a given dataset.

To evaluate the impact of the proposed conditionings, we apply them, by turn to the ground truths before computing the scores and estimating the distributions. We employ the Kolmogorov–Smirnoff (K–S) test to check which pairs of distributions are significantly different.

4 Results

The distributions of the Kappa scores observed, for the original ground truths, and for all proposed conditionings, appear in Figure 3.4 for the ISIC Archive, Figure 3.5 for the subset used in the ISIC 2017 Challenge and Figure 3.6 for the subset used in the ISIC 2018 Challenge.

The upper and lower parts of the figures plot the same information in a different form. The bottom part is perhaps more straightforward to interpret: shaded areas are the (normalized) histograms of the observed Cohen Kappa scores, and the line plots superimposed to them are the distributions estimated with a kernel density estimation. The upper part is more challenging to interpret but has the advantage to be much less crowded. In it, each experiment appears separated. The black dots represent the actual observations (with a small random horizontal jitter to help the visualization). The shapes around each group of points (violin plots) are the distributions estimated with a kernel density estimation (the shapes are more crowded where the

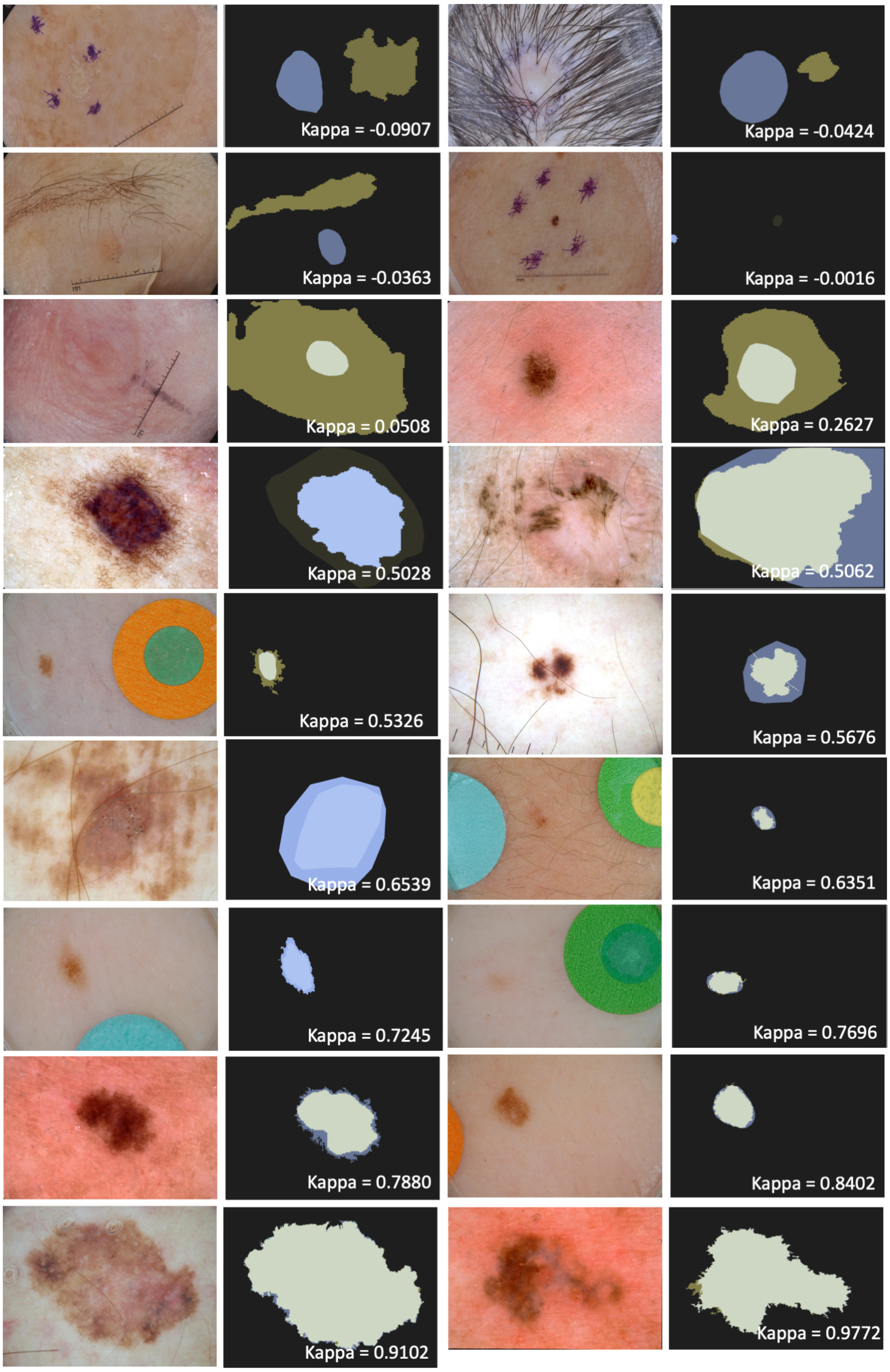

Figure 3.3: We have samples representative of all inter-annotator agreements we found in the distributions we observed. In our study, each skin lesion has at least two segmentation ground truths. The inter-annotator agreement is worse than random when the Kappa score is below 0. Kappa scores above 0.8 are considered high. The examples here may help the reader to appreciate the meaning of different scores qualitatively.

Percentile	ISIC 2017	ISIC 2018	ISIC Archive
25%	0.5724	0.5991	0.5748
50%	0.7438	0.7552	0.7185
75%	0.8213	0.8312	0.8010
95%	0.8838	0.8952	0.8812

Table 3.2: Percentiles of the Cohen's Kappa mean score distributions for each dataset.

distributions are denser), and the large red dots are the means of the distributions. The values of the scores for the quantiles of the original distributions are in Table 3.2 presents statistics for each dataset.

The distributions are highly skewed, with a robust mode towards high scores but a very long tail towards low scores. The most exciting result is that all conditionings improved the "good" mode considerably and that most of them are indistinguishable from each other in terms of that improvement. That is surprising since the morphological conditionings (opening and closing) are much more conservative than the convex hull, but all treatments combining those three operations obtained essentially the same results. Also surprising was that use of the bounding box — a much more destructive choice — was slightly *worse* than the other options.

None of the methods was able to improve the very divergent cases at the tail of the distribution: that was not unexpected since the small adjustments they make are not meant to reconcile those extreme cases. On the other hand, except for the bounding box, the techniques neither worsened the tail, which was a good outcome.

There is a small difference between the application of the convex hull and the use of the morphological operators alone, but we could not show that this difference is statistically significant. The K–S test rejected the equivalence of the original distribution with all conditionings, with tiny p-values (all p-values $< 10^{-20}$). It failed to reject most of the other pairs, with the notable exception of the bounding box vs. all conditionings with the convex hull ($10^{-7} <$ p-value < 0.002).

5 Discussion

Image segmentation is among the areas of computer vision that most advanced in recent years. Not only the techniques have improved sharply, but our understanding of the role of segmentation in the recognition pipeline, as well as its relationship with the task of classification, have changed drastically. However, obtaining properly annotated data to train and evaluate segmentation models continues to be a challenge. While datasets for (general) image classification have now millions of samples and thousands of categories, segmentation datasets are considerably smaller. For medical images, annotated data for segmentation is even scarcer.

Our results demonstrate the challenge of annotating skin lesions, showing that the median inter-annotation agreement for humans has around 0.72 Kappa score for the whole ISIC Archive and slightly more than that (0.75) for the images selected for the challenges. The good news is that straightforward image-processing techniques may significantly improve those agreements, without modifying too much the ground truths. Different applications may choose different conditionings according to their use cases: for location or demarcation, the convex hull may be the best, while for description the morphological operators, which preserve most of the border characteristics may be the best. An exciting result we found is that the substantial simplification brought by the bounding boxes worsened the annotation agreements in comparison to the other techniques.

From a theoretical point of view, we may interpret the conditioning as denoising operations, whose aim is to preserve the cogent information about the lesion segmentation, while discarding details which depend on the choice of one particular annotator. Therefore they may help both to train more robust machine learning models and to evaluate them more fairly.

The bad news is that none of the conditioning can deal with sharp divergences. Our results show that, although most masks have a reasonable-to-good inter-annotator agreement, there is a non-negligible tail of very disparaging annotations both in the ISIC Archive as a whole and on the subsets used on the challenges. That tail, and the difficulty in deciding which of the alternative annotations is the right one might explain why during the most recent challenge of 2018, none of the five top-ranked participants of the lesion boundary segmentation employed extra data for training (from the Archive, for example), while the four top-ranked participants for lesion classification (diagnosis) employed extra data.

In the next chapter, we will discuss our work on the evaluation of how our conditionings impact the research on machine learning models, by attempting to measure their effect on the training and evaluation of those models. Such evaluation is far from evident since the aim is to evaluate how models trained in a given setting generalize when exposed to different situations, in order to evaluate their robustness. As we will present, our design employs a cross-dataset evaluation, testing the models with images acquired and annotated under new conditions. An alternative design would be to use data augmentation techniques to simulate that design.

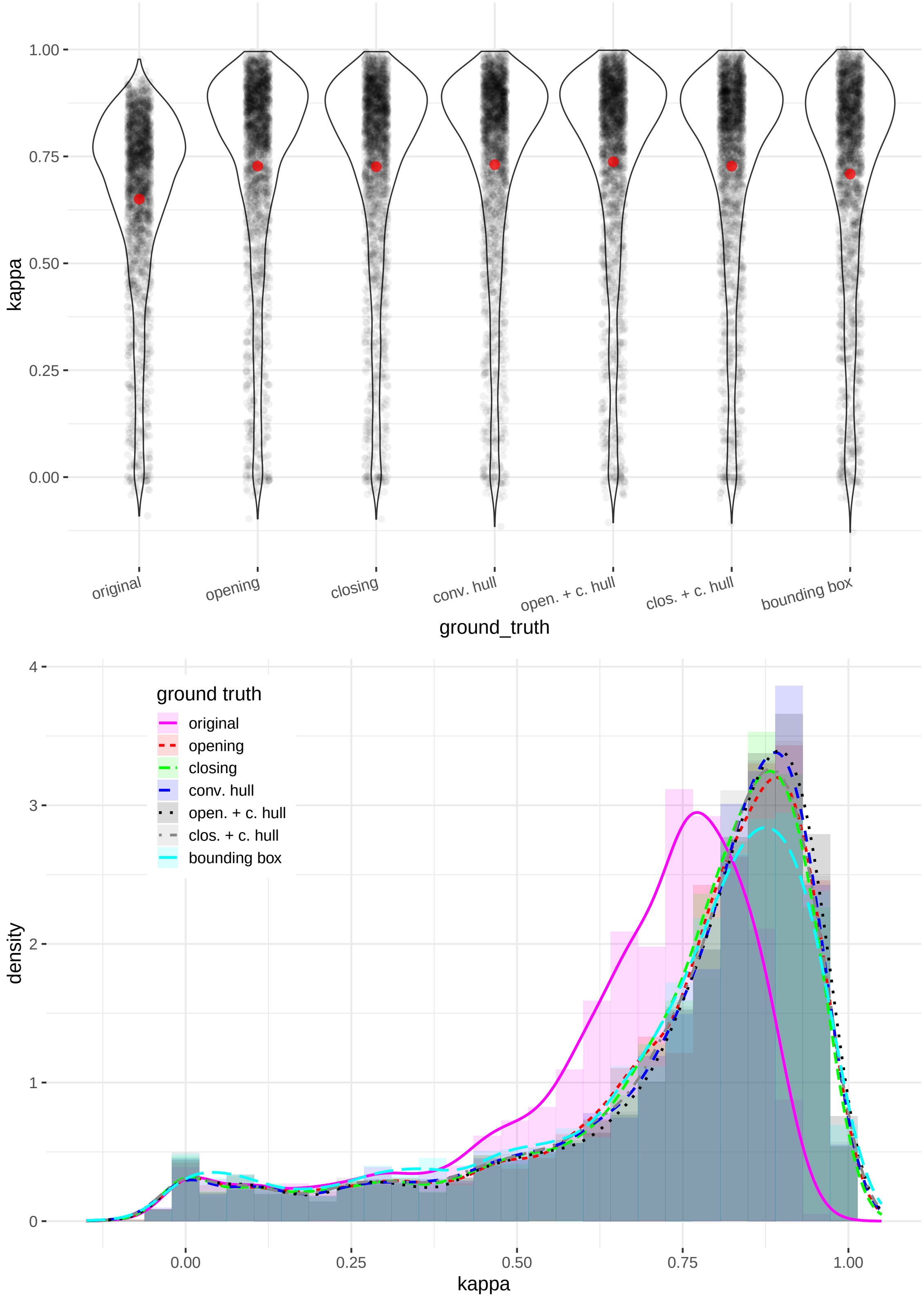

Figure 3.4: The distributions of inter-annotator agreements for the ground truths pre- (original) and post- the proposed conditionings (others). Both plots show precisely the same data. The bottom graph has the histograms (shaded areas) and the estimated densities (superimposed lines). The top graph has the original samples (black dots), the estimated densities (violin plots), and the estimated means (red dot) for each distribution. The plots show the data for the ISIC Archive.

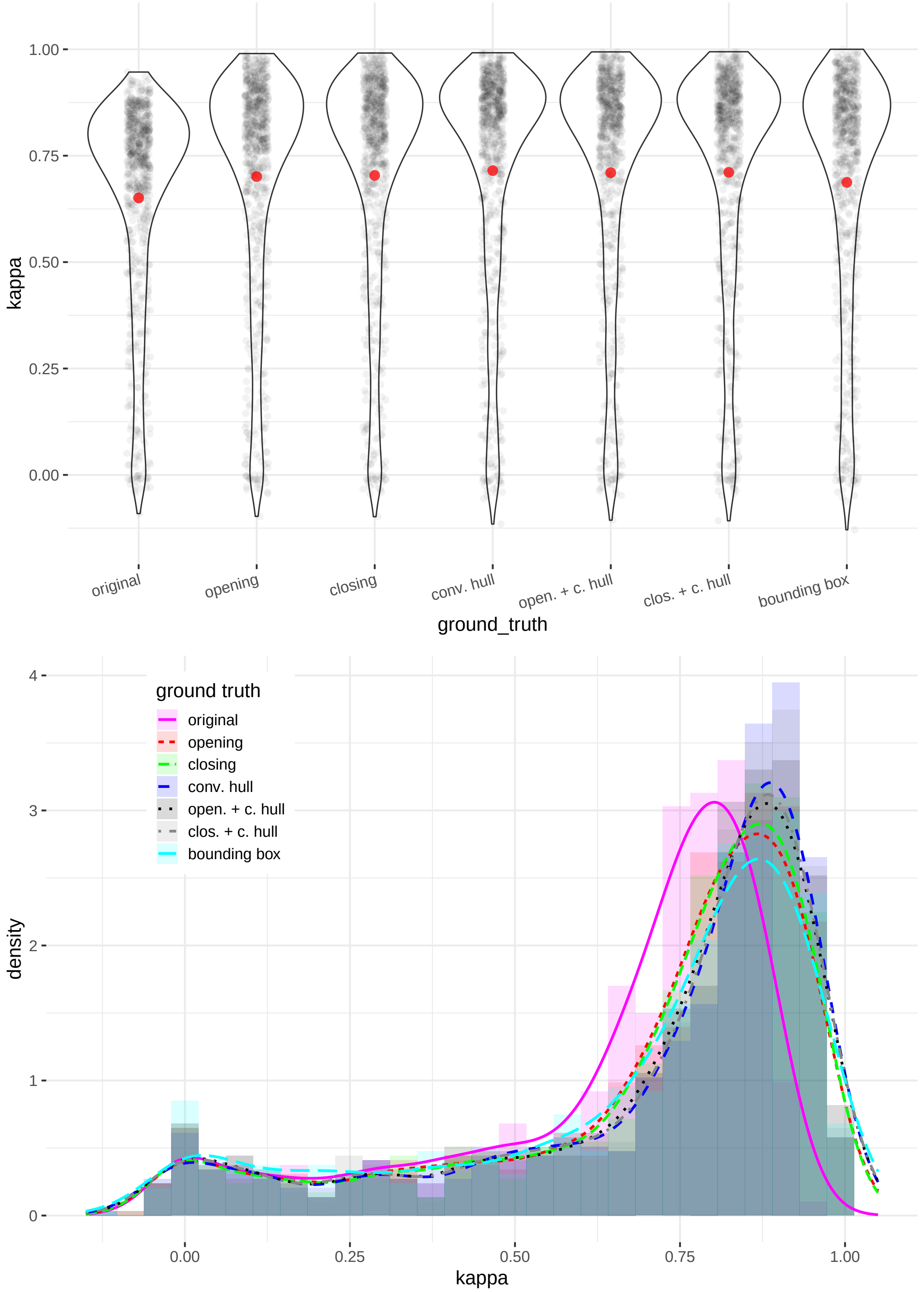

Figure 3.5: The distributions of inter-annotator agreements for the ground truths of the ISIC Challenge 2017. Please see Figure 3.4 and Section 4 for an explanation. Note how all proposed conditionings allow improving the inter-annotator agreement both here and on Figure 3.5.

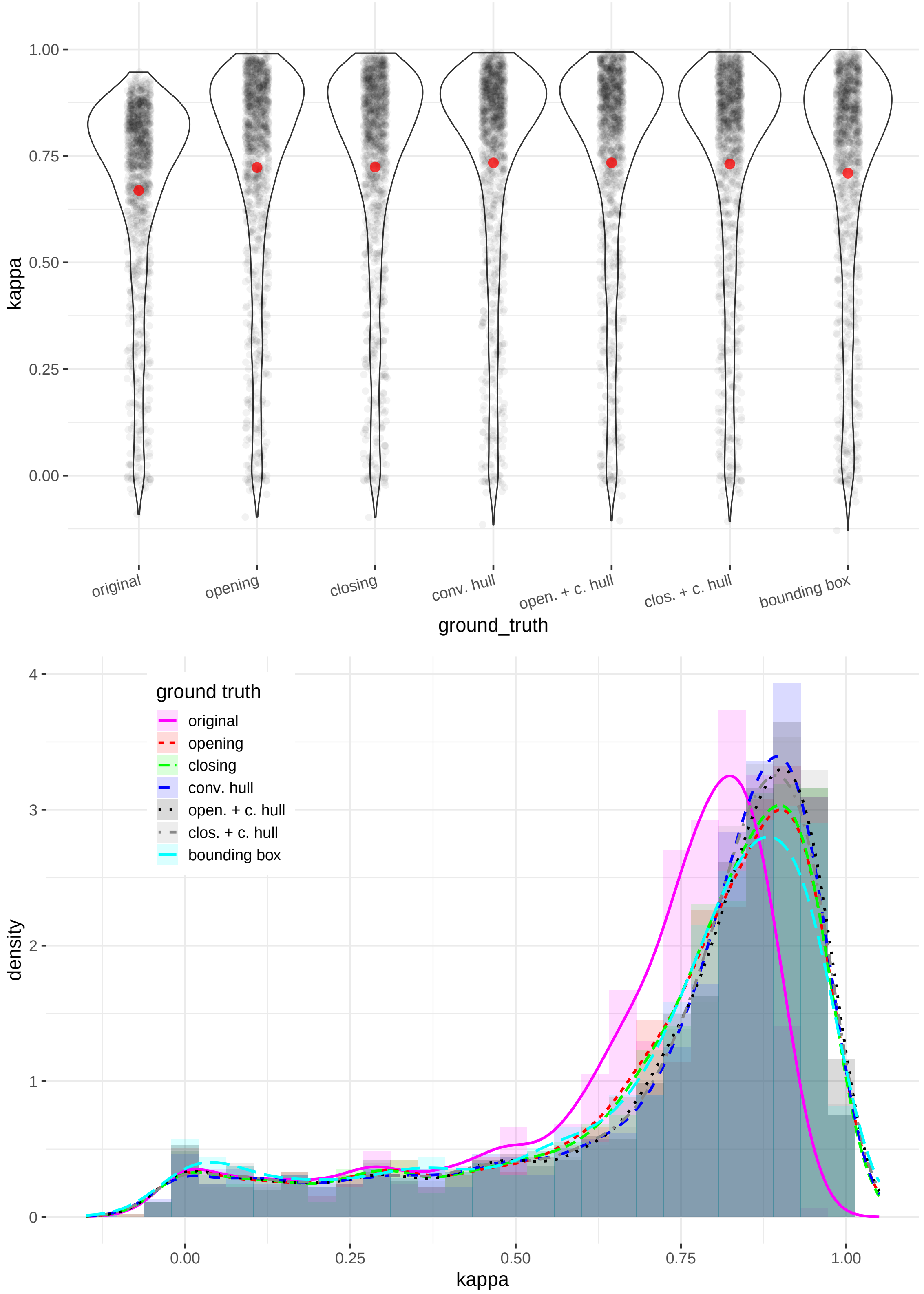

Figure 3.6: The distributions of inter-annotator agreements for the ground truths of the ISIC Challenge 2018. Please see Figure 3.4 and Section 4 for an explanation. Note how all proposed conditionings allow improving the inter-annotator agreement both here and on Figure 3.6.

Chapter 4

Skin Lesion Segmentation Under the Light of Inter-Annotator Agreement

In the previous chapter, we saw how the inter-annotator agreement strongly varies over the different annotations available for the most popular public dataset for skin lesion segmentation. As discussed before, the low agreement found can be a proxy for different inner characteristics of the data. It might suggest that the task is indeed complex. Many images can be ambiguous, and there is no single segmentation mask for a given lesion. These make the task arduous even for trained humans. On the other hand, a low agreement between several annotators can also be a proxy of the reliability of the dataset, and scientists should take it into account when evaluating the trained models.

With the previously described experiments, we saw how our conditioning could significantly increase the inter-annotator agreement between the masks of the ISIC Archive dataset. We could derive some conclusions from the analyzed data. First, there is a corpus of images with such a sharp disagreement that no treatment can reduce the gap between the annotations. Second, for the lesions with a higher agreement, all the proposed conditionings provide a similar impact.

In this section, we aim to evaluate the impact of our conditionings when dealing with machine learning models. We propose an experimental design with which we analyze how the training and the evaluation of the models vary with different factors considered. Then, we evaluate the models in a cross-dataset fashion.

Formally, the hypothesis we want to test in our experiments are:

- The conditioned models would *perform worse* than the non-conditioned models when *tested with the same dataset*;

- The conditioned models would *perform better* than the non-conditioned models when *tested with a different dataset*, i.e.they would generalize better in unknown scenarios;
- The observed effects are similar for all the proposed conditionings;
- Removing the tail of the distribution, i.e.the cases with a very low agreement, during training have no effect in the overall performance.

We organize the current section in the following manner. Section 1 describes the data used for the experiments. We explain the decisions made when designing the splits, how the data is distributed for training and validation, and give a brief overview of the datasets used for testing. Section 2 describes our experimental design. We discuss the decisions made for training and evaluation of the models, as well as the training scheme. Finally, in Section 3, we present our results, and we discuss our conclusions.

1 Dataset

We are recapping Table 3.1, which contains the distribution of the number of annotations per image for ISIC Archive, ISIC 2017 and ISIC 2018. We based the training of models for the current experiment in the subset of the ISIC Archive that has more than one annotation per lesion image. It is trivial but crucial reminder once the Cohen's Kappa score is only defined when there are *two* annotations for the same object. When there are more than two annotations, we calculated the score between all the two-by-two combinations of masks and took the mean between them.

From this subset, we derive two groups of data. The first contains all of the 2 233 lesions with two or more annotations. We call it *ISIC Full* from now on. The second contains the images with a Cohen's Kappa score higher than 0.5, a total of 1 808 lesion images. We call it *ISIC Clean* from now on. We split each of the datasets into training (80%) and validation (20%). The available images for each split is described in Table 4.1.

Split	ISIC Full	ISIC Clean
Training	1 786	1 449
Validation	447	359
Total	2 233	1 808

Table 4.1: Training and validation distributions for ISIC Full and ISIC Clean.

We want to make the final evaluation in a cross-dataset manner in order to challenge the generalization capabilities of our models when submitted to never-seen data. To accomplish this, we have three test sets. We built the first by randomly selecting 2 000 lesions from the remaining 11 546 images of the ISIC Archive that have only one segmentation mask. We call it *ISIC Titans* from now on. We decided to test with the subset of images with single annotation because we do not have any proxy of the reliability of the ground truth. This unknown scenario fits the idea of challenging the models to generalize when we have no control of the data. The two remaining test datasets are PH2 [64], a publicly available dataset with 200 dermoscopic annotated images, and the Edinburgh Dermofit Library [1], a private dataset with 1 300 dermoscopic annotated images.

2 Experimental Design

The first step on designing our experimental setup was to define the model architecture. Our baseline for experimentation is the ISIC 2018 competition. The competition organizers opened the system for late submissions for evaluation and created a Live Leaderboard [1] for ranking these submissions, which constitutes the perfect environment for robust experimentation. All submissions made to the Live Leaderboard were evaluated using the same criteria used for the ISIC 2018 Challenge [32].

For architecture and training configuration selection, we experimented with state of the art architectures strongly present in the literature. The models we experimented with were DeepLab V3+ [27], U-Net [77], LinkNet [26], AutoDeeplab [57] and RefineNet [56]. For all setups, we applied the Cyclic Learning Rate [81] strategy that we used during the ISIC 2018 Challenge. During our explorations for the competition, this technique was one of the most effective in advancing our results.

All models were trained using three data augmentation methods. The first was the addition of Gaussian Noise, the second was lowering down the contrast, and the third was degrading the color of the input image. Contrarily to the general idea of image pre-processing, that usually tries to remove noise, enhance contrast and color characteristics, the data augmentation we proposed does precisely the opposite. We degrade the input in order to make our models more robust and allow it to generalize better. Real data are not clean; the light may not be perfect

[1] https://challenge2018.isic-archive.com/live-leaderboards/

and have low contrast between lesion and background healthy skin. Going against common sense and forcing the training data to hold these adverse conditions is, in fact, the right approach.

We trained all five architectures with the ISIC 2018 [4] training dataset and submitted them to evaluation under the validation set of the Live Leaderboard. Table 4.2 shows the best result for each trained model using the Jaccard Index with 0.65 threshold metric.

Rank	Model	Threshold Jaccard Index
4	DeepLab V3+	0.793
13	AutoDeeplab	0.762
20	LinkNet	0.749
28	RefineNet	0.731
33	U-Net	0.717

Table 4.2: The best result for each model when submitted to evaluation under the validation set of ISIC 2018 Live Leaderboard. Column *Rank* corresponds to the rank in the validation leaderboard at the time of the writing of this study.

After extensive experimentation, the model that excelled was DeepLab V3+, which is consistent with the State of the Art described in Section 4. We then submitted the model to a final evaluation in the test dataset. Table 4.3 presents the primary and secondary metrics available in the Live Leaderboard for the architecture.

Notice that our model is considerably below the top submission in the Live Leaderboard, which achieved the score of 0.832 in the primary metric until the moment of the writing. There are three things to consider about the results. The first is that the top submitters are not clear about the methods and training configuration they used to accomplish their results. The second is that many models used to reach top positions in competitions are ensembles, not single model predictions as we desire. With our experiments, we want to understand how applying conditionings to the models may affect the training and evaluation of single models, and extending these ideas to ensembles is out of the scope of the study. Finally, the third thing is that we complied with all of the good practices of machine learning development. For scientific methods, it is not reasonable to try unexplainable things *just because* they seem to have a positive effect. During experimentations, we did not try anything without proper reasoning, which makes our results robust and reproducible. With these considerations, we understood it would be prudent to accept the presented results as a good baseline.

With a solid decision about which architecture to proceed with, we follow on with the experimental design to evaluate the impact of conditioning on the training of machine learning

DeepLab V3+ Best Results	
Metric	Score
Threshold Jaccard Index*	0.750
Jaccard Index	0.807
Dice Coefficient	0.883
Sensitivity	0.938
Specificity	0.921
Accuracy	0.934

Table 4.3: The best result for DeepLab V3+ when submitted to evaluation under the test set of ISIC 2018 Live Leaderboard. Column *Threshold Jaccard Index*, marked with a *, is the primary metric for the competition.

models. The next decision is which conditionings to test. As described in the previous section, all of them have positive and similar effects on the inter-annotator agreement statistics. For this reason, we decided to work with only two of them: morphological opening and convex hull. Both of them have a simple implementation and have completely different visual effects. While the opening operation maintains the general look of the lesion, destroying only the method-specific border characteristics, the convex hull is much more destructive, building a final segmentation mask more similar to the polygon method.

As described in Section 1, we ran the same experiments on both ISIC Full and ISIC Clean. By design, the lesion images on both datasets have a minimum of two annotations. During training, for each epoch, we randomly select one of the available annotations to use as ground truth. This design gives robustness to the model once it can learn that there is more than one possible right answer to the problem, and it helps to avoid overfit. During validation, we evaluate the model on all available annotations, and we use as the real metric the best one, i.e., highest Jaccard Index. All of the test datasets have single annotation per lesion, so mask selection is not an issue during this phase.

For training and testing, there are three possible settings: no conditioning, opening, and convex hull. During the validation phase, we did not apply conditionings to the segmentation masks. We selected the final model by choosing the one with the best performance during validation. We ran the described experimental design for both ISIC Full and ISIC Clean, resulting in six trained models.

Notice that the proposed experiment design, as well as the design of the datasets, are entirely aligned with the hypothesis we want to validate. For the first and the second hypothesis, we use multiple testing datasets. The ISIC Titans dataset has images extracted from the same distribution as the images in the training datasets. The PH2 and the Edinburg Dermofit are composed of images taken from unknown distributions. For the third hypothesis, we selected two conditionings with very different effects. Finally, to validate the last hypothesis, we have two different training datasets, the ISIC Full, and the ISIC Clean.

The code for the experiments described in this Chapter is public available at our Github repository [2].

3 Results and Discussion

We ran each of the described experiments three times for the sake of removing the random variability present during training. We ended up with six trained configurations, and we tested each of them under the three conditionings proposed — no conditioning, opening, and convex hull — for the three test datasets — ISIC Titans, Edinburgh Dermofit and PH2. The reported metrics are the same used for the ISIC 2018 Challenge — Jaccard Index, i.e., intersection over union, and Threshold Jaccard Index, i.e.if the Jaccard Index is below a particular value (0.65), the prediction receives 0.0 score.

The results are presented as follows: Table 4.4, Table 4.5 and Table 4.6 compare different conditioning types, for different training sets in ISIC Full and ISIC Clean when evaluating on *ISIC Titans*, *PH2*, and *Edinburgh Dermofit*, respectively. The results columns present the mean Jaccard Index and the mean Jaccard Index with 0.65 threshold for each testing set. All the experiments have a standard deviation, which is not present in the tables, between different runs of the same experiment. For the complete results of our experiment, we refer the reader to Appendix A.

Figures 4.1 and 4.2 present interaction plots between the the experimented factors. The first plots the interactions between the two train dataset, the conditionings during the train, and the three test datasets. The factor *conditioning during the test* is set as no conditioning. The second plots the interactions between the conditionings during the train, the conditionings during the test, and the test datasets. The factor *training dataset* is set as ISIC Clean.

The structure of the chart is: The colored dots are the results of each of our individual experiments. The solid colored lines plot the mean of each experiment. The dashed black line plot the mean of all the experiments in the column. Each color represents one testing dataset.

[2] https://github.com/vribeiro1/inter-annotator-agreement-skin-lesion-segmentation

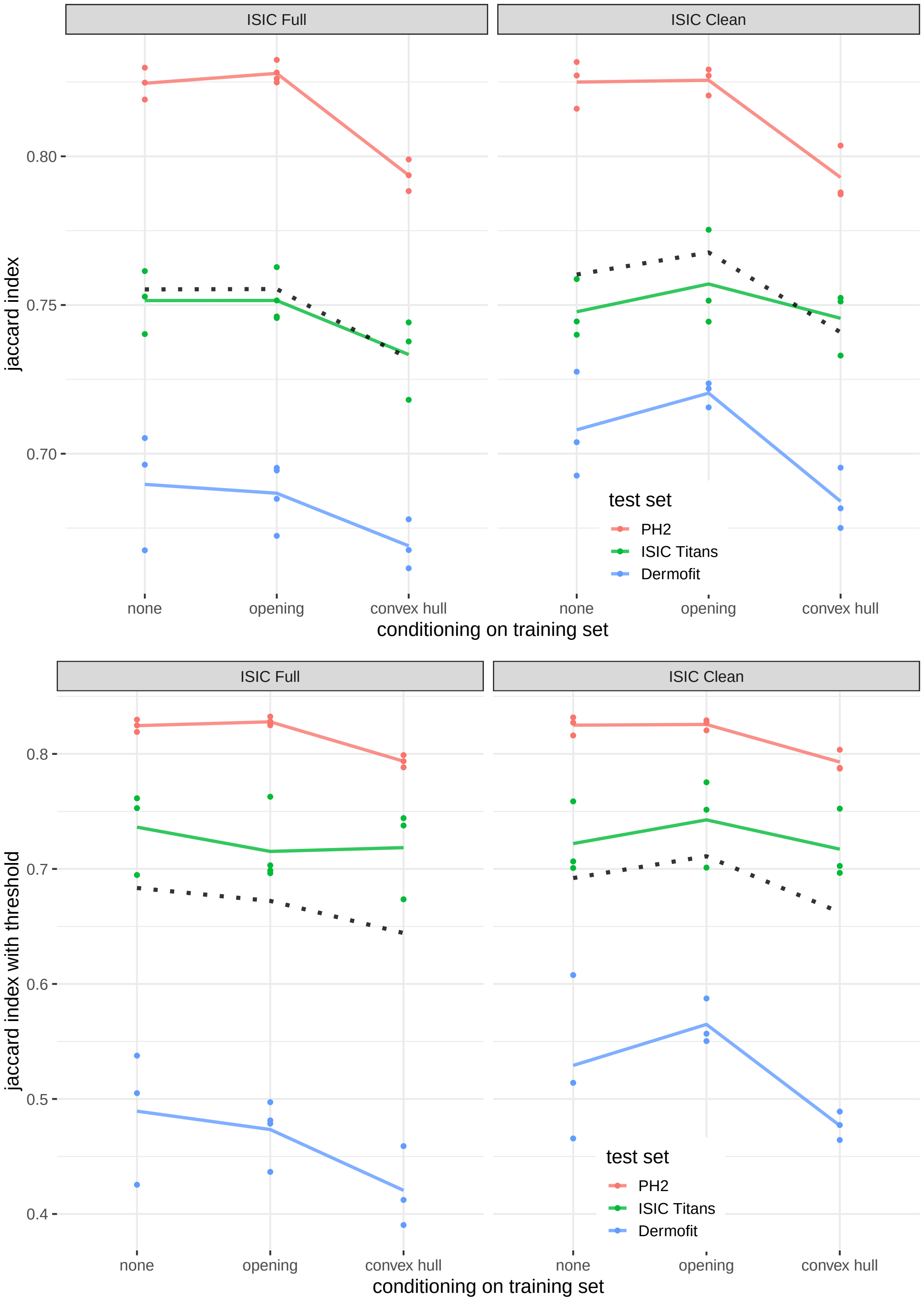

Figure 4.1: Interaction plots between the factors in the experiment fixating **no conditioning during test**.

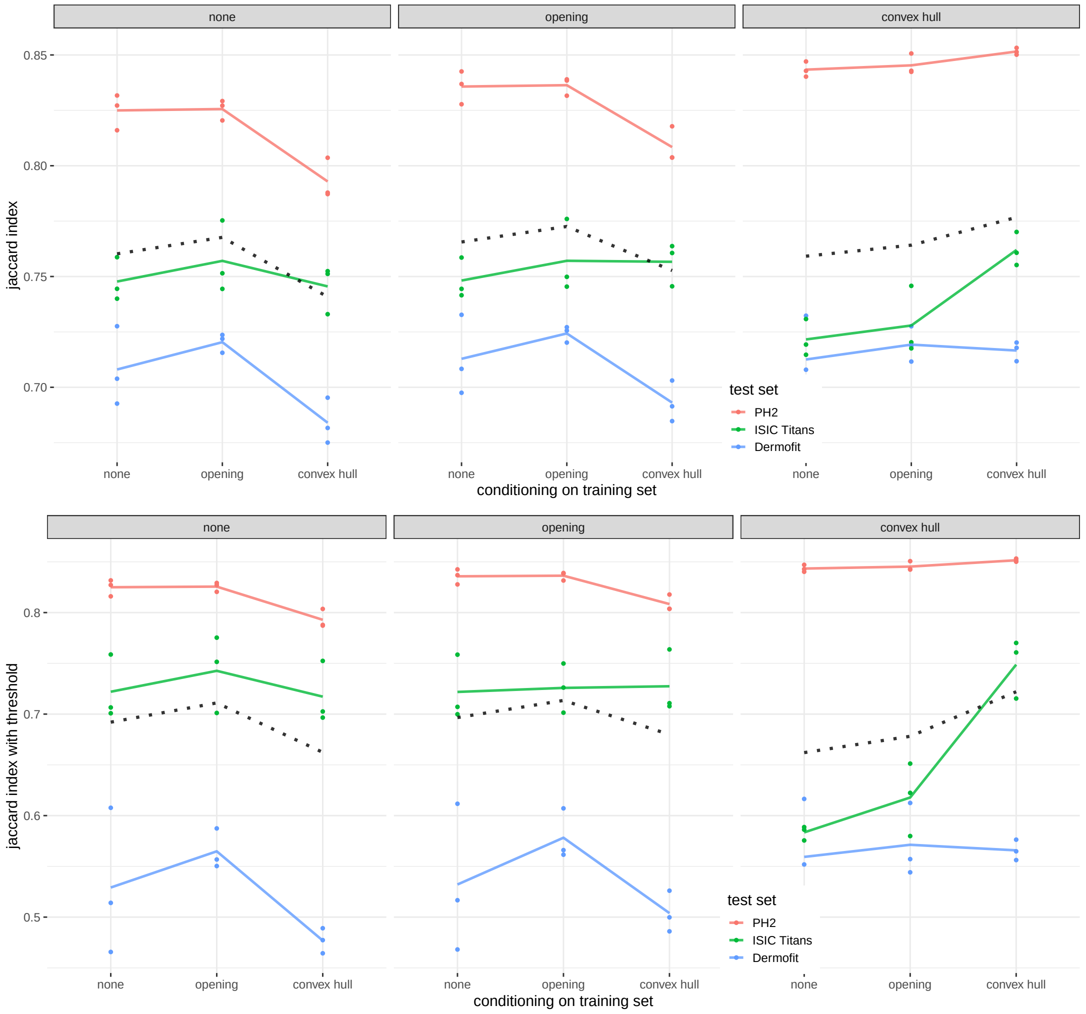

Figure 4.2: Interaction plots between the factors in the experiment fixating **ISIC Clean as training set**.

From Figure 4.1 we can observe by the dashed black line that the removal of the images with the low inter-annotator agreement is beneficial for the performance of the model. This result is non-trivial once our cutting point ($\kappa < 0.5$) removes the long tail of Cohen's Kappa distribution, and it represents close to 19% of the training data. For deep learning, reaching *similar* metrics with less data is already hard enough, even hard is to outperform.

Another interesting result is the observed improvement of the results when we apply the opening operation. This conditioning destroys the details in the border of the lesion in the annotation mask. We train the models without those details, but when we test with the original detailed mask, the model is still able to reconstruct them and perform better than the models trained with non-conditioned masks. On the other hand, when we train the models with the convex hull, a very destructive operation, the result strongly deteriorates. In this case, because we removed too much information from the training data, the model is not able to reconstruct the details during the testing phase.

From Figure 4.2, when we fix the training dataset as ISIC Clean and add the factor *conditioning during testing*, we see that when we test the models with the convex hull, the model that excels in the one trained with the convex masks. In this case, when we do not care about the border details, and we only want the envelope of the lesion, it is indeed better to train our models with the convex annotations.

We ran an analysis of variance (ANOVA) for the whole experiment, which suggests that all of the main effects are strongly significant. For a complete view of our ANOVA and the significance of the experiments, we refer the reader to Appendix B.

From these results, we can derive some conclusions. The ISIC Full has a substantial disagreement between the annotations, and more importantly, it has a large corpus of ambiguous and complex images, which have a close-to-zero agreement (sometimes less). Because of these characteristics, the conditionings are not able to solve the thick tail of the distribution, and the networks cannot take advantage of the positive impacts generated by modifying the masks. When we remove the ambiguity, the applied transformations work as a regularizer for the inputs, removing the method-specific characteristics of the ground truths and leaving only the information that is useful for the segmentation task.

When comparing the two proposed conditionings, it is clear how the opening conditioning drives better results than the convex hull when we require some level of details. We can observe it for both training sets. The only case where the former conditioning surpasses the first is when the target is also convex, i.e.when we apply convex hull during testing. In the neutral case, i.e.when we test with the original ground truths, the opening operation is consistently better. This result goes in the opposite direction of the ones presented in Chapter 3. Looking at the isolated inter-annotator agreement metrics, we expected the conditioning to be fully equivalent. Both improved the inter-annotator agreement with no statistically significant difference between each other. However, submitting the operations to the challenge of improving segmentation metrics, we understand that they are indeed different, and some conditionings may be better than others for specific use cases.

On the one hand, the opening operation is more conservative and preserve most of the information present in the ground truth mask. On the other hand, the convex hull is more destructive. As described in Chapter 3, the opening conditioning would be more appropriate as a pre-processing operation for a description task, when we need to characterize the lesion fully. However, when we care about localization — when we need a rough estimation of where the lesion is in the patient's body — generating a convex annotation around the lesion demands less human effort than a fine one.

Summarizing, we started with our research with four hypotheses. From our experiments, we can see that three of them are not valid, while only one is valid.

The first states that the conditioned models would *perform worse* when tested with the same data, which is not true. When we tested with data extracted from the same distribution of the training dataset, we observed that condioned models improved the results. The second states that the conditioned models would *perform better* when tested with unknown data, which is valid. Our results show that the models trained with the opening conditioning have better performance than the non-conditioned models. The third hypothesis states that the different conditionings would have a similar effect since we did not observe any statistical difference between conditionings in Chapter 3. This hypothesis is invalid. When we require details in the testing phase, the opening conditioning is better, while when the target mask is convex, it is better to train with the convex annotation. Finally, our last hypothesis states that the removal of the images with low Cohen's Kappa Score would not affect the model performance, which is invalid. The models trained with ISIC Clean show significantly better results than those trained with ISIC Full.

The results presented in this section is auspicious, and it introduces a discussion commonly ignored by the scientists. The skin lesion community should think less in the isolated metrics for skin lesion segmentation, which in many cases are saturated and very close to the human level, and spend more efforts thinking what are the final goals and use cases we aim. To the best of our knowledge, this is the first work to raise the inter-annotator agreement discussion in the skin lesion segmentation area, especially proposing methods to reduce disagreement and to compare the performances when training machine learning models, and we are sure there is much more to explore in this subject.

		Cond. Test - Jacc.			Cond. Test - Jacc. Thres.		
Dataset Train	Cond. Train	None	Opening	Convex Hull	None	Opening	Convex Hull
ISIC Full	None	0.751	0.754	0.732	0.736	0.739	0.622
	Opening	0.752	0.753	0.730	0.715	0.729	0.623
	Convex Hull	0.733	0.746	0.754	0.718	0.730	0.723
ISIC Clean	None	0.748	0.748	0.722	0.722	0.722	0.583
	Opening	0.757	0.757	0.728	0.743	0.726	0.618
	Convex Hull	0.746	0.757	0.762	0.717	0.727	0.749

Table 4.4: Comparison between different conditionings types and testing sets, for training in ISIC Full and ISIC Clean, when evaluating in the **ISIC Titans** dataset. Note: All experiments a presented standard deviation between the different runs of the same experiment.

		Cond. Test - Jacc.			Cond. Test - Jacc. Thres.		
Dataset Train	Cond. Train	None	Opening	Convex Hull	None	Opening	Convex Hull
ISIC Full	None	0.825	0.836	0.850	0.825	0.836	0.850
	Opening	0.828	0.839	0.851	0.828	0.839	0.851
	Convex Hull	0.794	0.810	0.855	0.794	0.810	0.855
ISIC Clean	None	0.825	0.836	0.843	0.825	0.836	0.843
	Opening	0.826	0.836	0.845	0.826	0.836	0.845
	Convex Hull	0.793	0.808	0.852	0.793	0.808	0.852

Table 4.5: Comparison between different conditionings types and testing sets, for training in ISIC Full and ISIC Clean, when evaluating in the **PH2** dataset. Note: All experiments presented a standard deviation between the different runs of the same experiment.

		Cond. Test - Jacc.			Cond. Test - Jacc. Thres.		
Dataset Train	Cond. Train	None	Opening	Convex Hull	None	Opening	Convex Hull
ISIC Full	None	0.733	0.794	0.669	0.718	0.794	0.421
	Opening	0.746	0.810	0.678	0.730	0.810	0.436
	Convex Hull	0.754	0.855	0.700	0.723	0.855	0.510
ISIC Clean	None	0.690	0.695	0.698	0.489	0.506	0.520
	Opening	0.687	0.692	0.695	0.473	0.482	0.505
	Convex Hull	0.669	0.678	0.700	0.421	0.436	0.510

Table 4.6: Comparison between different conditionings types and testing sets, for training in ISIC Full and ISIC Clean, when evaluating in the **Dermofit** dataset. Note: All experiments presented a standard deviation between the different runs of the same experiment.

Chapter 5

Conclusions

On this work, we reviewed the art of skin lesion segmentation since the pre-deep learning times until the most recent generative models. We discussed the many works submitted to the 2017 and 2018 editions of the ISIC Challenge: Skin Lesion Analysis Towards Melanoma Detection. We also described our submission (RECOD Titans) to the segmentation task and showed all the challenges we faced along with the competition.

When reviewing the state of our research, we found many unexplored paths in the research field. We thoroughly explored the ISIC Archive and its subsets for each edition of the challenge under the light of annotator agreement. The analysis generated an article [76] that is currently in its pre-print version, and we aim to publish it as soon as possible in a high impact journal.

Finally, we developed an experimental design to evaluate our proposed conditionings when training and evaluating machine learning models. The results of the experiments are fascinating as they show that it is reasonable to be more careful about the data we use for semantic segmentation since the significant disagreement between annotators might expose ambiguity intrinsic to the data and deteriorate the model's capability of generalization in a cross-dataset fashion. Also, our experiments raise a discussion often left apart by the scientific community, which is the one about the use cases of semantic segmentation.

The discussions we proposed are very novel and raise essential topics when we think about bringing the knowledge achieved in the academic field to society. Addressing melanoma detection outside the laboratory is not trivial. It requires understanding the data, the use cases, the daily life of physicians and other health professionals, and most importantly, understanding the impact that wrong predictions may bring to patients' lives. We expect that the analysis we raised about the inter-annotator agreement distribution present in an essential dataset for skin lesion segmentation enables more discussion about how the data available for the medical field impacts predictions and what we can do to overcome the issues.

1 Contributions

Along with these studies, we could develop a solid knowledge about the art of skin lesion segmentation. Our research group is very experienced with machine learning tools for skin lesion classification. We have already five years of experience working in the medical field, especially with melanoma, and during these years, we achieved many goals and earned many prizes. Not only on the deep learning theory and practice, but the group's work also have a research line discussing how to get the work developed inside the laboratory to the real work.

Although the research on *classification* is very advanced and achieving great results, we are just starting in skin lesion *segmentation*. This thesis is just the second we have for the task. The author's contributions during the development of this work are:

- In-deep exploration of the current state of the art of skin lesion segmentation. We discussed the art since the pre-deep learning times, the current art for medical images and the last two editions of the ISIC Challenge.
- Participation in the ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection.
- Being the first work, to the best of our knowledge, to introduce the discussion of the inter-annotator agreement for skin lesion segmentation. We not only analyze the data for the largest public dataset in the field, but we proposed methods for improving the metrics and explored the impact of the conditioning when training and evaluating machine learning models.

2 Achievements

During the M. Sc. program, the author also had relevant achievements that deserve a highlight. The author contributed the following two scientific publications:

- Deep-Learning Ensembles for Skin-Lesion Segmentation, Analysis, Classification: RECOD Titans at ISIC Challenge 2018 [19].
- Handling Inter-Annotator Agreement for Automated Skin Lesion Segmentation [76].

The author also enrolled in the Summer School in Data Science for Document Analysis and Understanding offered by the University of La Rochelle & INRIA, France, July 2019. The participation in the program was funded by the French Embassy in Brazil, which provided the flight tickets, and by Nexa Digital, which funded the course's fee.

Finally, it is relevant to highlight that all the program was self-funded, and during the studies, the author worked as both a data scientist and software engineer at different well-recognized companies in many different areas. Working in the software industry provided contact with the most modern tools and techniques used by the companies to deal with real-world problems. Although very energy consuming, working at the industry and the university concurrently was essential to develop in-depth and reliable knowledge in data science and engineering.

3 Future Work

Our experiments suggest that the existing datasets for skin lesion segmentation have ambiguous images, which complicates the learning of artificial intelligence methods. ISIC Archive is the only public dataset that contains multiple annotations for each image, and even for this one, the multi-annotated lesions are a minority. Moreover, the single-annotated images give no proxy about the reliability of the ground truth and the ambiguity of the lesion.

From the data view, our work proposes a framework for training the models with multiple annotations per image. We randomly select one of the available segmentation masks at each epoch. This approach helps the model to learn that there is no unique truth about the borders of the lesion.

Although straightforward, the proposed pipeline does not entirely fit the case where we have one annotation for each image. Understanding how to incorporate this set during training can be valuable since we sharply increase the number of available data points. For lesions with a lower number of annotations, we have a lower probability of disagreement between annotators, but we have a smaller number of evidence about the exact borders of the lesion.

On that matter, one approach that we can explore is STAPLE [87], which presents an expectation-maximization algorithm for simultaneous truth and performance level estimation. In the author's words, the algorithm considers a collection of segmentations and computes a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation. Applying STAPLE can be a good alternative to the annotation selection process and help to incorporate both single and multi-annotated images.

From the model view, our framework only fits a single model evaluation. It is out of our scope to extend these experiments to ensembles. However, training multiple models is common in many deep learning pipelines. Ensemble models work as multiple opinions of experts about the nature of the problem. An extension of our proposal is to evaluate *if*, and *how* ensembles take advantage of the conditionings during learning and compare with the single models.

Another reasonable research line would be exploring multi-objective learning for skin lesion segmentation. Detecting the borders of the lesion is not the only interest with segmentation. A related task is to segment medical attributes in the lesion like pigment network, negative network, globules, and others. Exploring a network capable of learning both tasks together and sharing information between them can make the model outperform both tasks when compared to single-objective models.

During our experiments, we did not evaluate how segmentation and the proposed conditioning approaches fit the classification of skin lesions. The work of our research group suggests that incorporating segmentation as a feature of classification is not trivial. Future work may evaluate how our conditionings affect this result and if they could be used to address better mask selection processes.

A similar task to segmentation, but incorporating the idea of skin lesion classification would be understanding how to generate predictions about each pixel of the image, but instead of predicting if they belong to the lesion or not, we can predict how they affect the overall diagnosis of the lesion. Which pixels explain the choice for the overall class of the image and which of them do not. Addressing accountability for machine learning is an active and exciting research area.

Future work also includes exploring the uncertainty on the skin lesion ground truths and training the machine learning methods to be robust to ambiguity and disagreement between annotators. The work of Aroyo and Welty [14] argues that the disagreement is not noise; it brings rich information about the nature of the problem. Understanding how to use this signal for building more reliable systems may be fruitful.

Bibliography

[1] Dermofit image library. https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html.

[2] General data protection regulation. https://gdpr-info.eu/.

[3] Imagenet. URL http://www.image-net.org/.

[4] ISIC 2018 - Task 1: Lesion Boundary Detection, . https://challenge2018.isic-archive.com/task1/.

[5] International Skin Imaging Collaboration: Melanoma Project, . https://isic-archive.com.

[6] Lei geral de proteção de dados. http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm.

[7] NumPy. http://www.numpy.org/.

[8] A practical guide to autoencoders. https://sadanand-singh.github.io/posts/autoencoders/.

[9] Scikit-Image: Image processing in Python. https://scikit-image.org/.

[10] Deeplab image semantic segmentation network. https://sthalles.github.io/deep_segmentation_network/.

[11] Adegun Adekanmi Adeyinka and Serestina Viriri. Skin lesion images segmentation: A survey of the state-of-the-art. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 321–330. Springer, 2018.

[12] Vamshi Ambati, Stephan Vogel, and Jaime G Carbonell. Active learning and crowd-sourcing for machine translation. In *LREC*, volume 1, page 2, 2010.

[13] Muhammad Ammar, Sajid Gul Khawaja, Abeera Atif, Muhammad Usman Akram, and Muntaha Sakeena. Learning based segmentation of skin lesion from dermoscopic images. In *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE, 2018.

[14] Lora Aroyo and Chris Welty. Measuring crowd truth for medical relation extraction. In *2013 AAAI Fall Symposium Series*, 2013.

[15] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017.

[16] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees. 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. Springer, 2013.

[17] Matt Berseth. Isic 2017-skin lesion analysis towards melanoma detection. *arXiv preprint arXiv:1703.00523*, 2017.

[18] Lei Bi, Jinman Kim, Euijoon Ahn, and Dagan Feng. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. *arXiv preprint arXiv:1703.04197*, 2017.

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

[20] Alceu Bissoto, Fábio Perez, Eduardo Valle, and Sandra Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 294–302. Springer, 2018.

[21] T. J. Brinker, A. Hekler, A. Hauschild, C. Berking, B. Schilling, A. H. Enk, S. Haferkamp, A. Karoglan, C. von Kalle, M. Weichenthal, et al. Comparing artificial intelligence algorithms to 157 german dermatologists: the melanoma classification benchmark. *European Journal of Cancer*, 111:30–37, 2019.

[22] M Emre Celebi, Y Alp Aslandogan, and Paul R Bergstresser. Unsupervised border detection of skin lesion images. 2:123–128, 2005.

[23] M Emre Celebi, Hitoshi Iyatomi, Gerald Schaefer, and William V Stoecker. Lesion border detection in dermoscopy images. *Computerized medical imaging and graphics*, 33(2):148–153, 2009.

[24] M EMRE Celebi, QUAN Wen, HITOSHI Iyatomi, KOUHEI Shimizu, H Zhou, and G Schaefer. A state-of-the-art survey on lesion border detection in dermoscopy images. *Dermoscopy Image Analysis*, pages 97–129, 2015.

[25] S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman, and L. Tarassenko. Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 266–272, 2017.

[26] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

[27] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[28] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[29] Sheng Chen, Zhe Wang, Jianping Shi, Bin Liu, and Nenghai Yu. A multi-task framework with feature passing module for skin lesion classification and segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1126–1129. IEEE, 2018.

[30] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. pages 2843–2851, 2012.

[31] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, 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*, pages 168–172, 2018.

[32] N. C. F. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, 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.

[33] Marceli de Oliveira Santos. Estimativa 2018: incidência de câncer no brasil. *Revista Brasileira de Cancerologia*, 64(1):119–120, 2018.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[35] Yining Deng and BS Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE transactions on pattern analysis and machine intelligence*, 23(8):800–810, 2001.

[36] A. P. Dhawan and A. Sim. Segmentation of images of skin lesions using color and texture information of surface pigmentation. *Computerized Medical Imaging and Graphics*, 16(3):163–177, 1992.

[37] J. Egger, K. Hochegger, M. Gall, K. Reinbacher, K. Schwenzer-Zimmerer, J. Wallner, and D. Schmalstieg. Clinical evaluation of mandibular bone segmentation. *IEEE Engineering in Medicine and Biology Society*, 2016.

[38] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[40] M. Fornaciali, M. Carvalho, F. V. Bittencourt, S. Avila, and E. Valle. Towards automated melanoma screening: Proper computer vision & reliable results. *arXiv preprint arXiv:1604.04024*, 2016.

[41] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. pages 2672–2680, 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[42] Adele Green, Nicholas Martin, John Pfitzner, Michael O'Rourke, and Ngaire Knight. Computer image analysis in the diagnosis of melanoma. *Journal of the American Academy of Dermatology*, 31(6):958–964, 1994.

[43] Danna Gurari, Kun He, Bo Xiong, Jianming Zhang, Mehrnoosh Sameki, Suyog Dutt Jain, Stan Sclaroff, Margrit Betke, and Kristen Grauman. Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s). *International Journal of Computer Vision*, 126(7):714–730, 2018.

[44] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, 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.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[46] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[47] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun. Deep learning for image-based cancer detection and diagnosis- a survey. *Pattern Recognition*, 83:134–149, 2018.

[48] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[49] Charles Jones, Christopher Tonetti, et al. Nonrivalry and the economics of data. In *2018 Meeting Papers*, volume 477. Society for Economic Dynamics, 2018.

[50] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajeddin, Ali Gooya, and Nasir Rajpoot. Leveraging transfer learning for segmenting lesions and their attributes in dermoscopy images. *arXiv preprint arXiv:1809.10243*, 2018.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[52] Thomas Küstner, Sarah Müller, Marc Fischer, Jakob Weiβ, Konstantin Nikolaou, Fabian Bamberg, Bin Yang, Fritz Schick, and Sergios Gatidis. Semantic organ segmentation in 3d whole-body mr images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3498–3502. IEEE, 2018.

[53] T. A. Lampert, A. Stumpf, and P. Gançarski. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing*, 25(6):2557–2572, 2016.

[54] George Leifman, Tristan Swedish, Karin Roesch, and Ramesh Raskar. Leveraging the crowd for annotation of retinal images. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7736–7739. IEEE, 2015.

[55] M. Liedlgruber, K. Butz, Y. Höller, G. Kuchukhidze, A. Taylor, O. Tomasi, E. Trinka, and A. Uhl. Variability issues in automated hippocampal segmentation: A study on out-of-the-box software and multi-rater ground truth. In *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 191–196, June 2016. doi: 10.1109/CBMS.2016.55.

[56] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[57] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *arXiv preprint arXiv:1901.02985*, 2019.

[58] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[59] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[60] Oeslle Lucena, Roberto Souza, Letícia Rittner, Richard Frayne, and Roberto de Alencar Lotufo. Convolutional neural network for brain mr imaging extraction using silver-standards masks. 2018.

[61] M. A. Marchetti, N. C. F. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, 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.

[62] David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Iccv Vancouver:, 2001.

[63] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3): 276–282, 2012.

[64] T. F. Mendonca, M. E. Celebi, T. Mendonca, and J. S. Marques. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis*, 2015.

[65] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle. Knowledge transfer for melanoma screening with deep learning. In *IEEE International Symposium on Biomedical Imaging*, pages 297–300, 2017.

[66] Afonso Menegola, Julia Tavares, Michel Fornaciali, Lin Tzy Li, Sandra Avila, and Eduardo Valle. Recod titans at isic challenge 2017. *arXiv preprint arXiv:1703.04819*, 2017.

[67] Philippe Meyer, Vincent Noblet, Christophe Mazzara, and Alex Lallement. Survey on deep learning for radiotherapy. *Computers in biology and medicine*, 2018.

[68] R. H. Moss, G. A. Hance, S. E. Umbaugh, and W. V. Stoecker. Unsupervised color image segmentation: with application to skin tumor borders. 1996.

[69] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

[70] Samina Naz, Hammad Majeed, and Humayun Irshad. Image segmentation using fuzzy clustering: A survey. pages 181–186, 2010.

[71] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[72] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 303–311. Springer, 2018.

[73] Chengyao Qian, Ting Liu, Hao Jiang, Zhe Wang, Pengfei Wang, Mingxin Guan, and Biao Sun. A two-stage method for skin lesion analysis. *arXiv preprint arXiv:1809.03917*, 2018.

[74] Tran Minh Quan, David GC Hildebrand, and Won-Ki Jeong. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. *arXiv preprint arXiv:1612.05360*, 2016.

[75] MI Rajab, MS Woolfson, and SP Morgan. Application of region-based segmentation and neural network edge detection to skin lesions. *Computerized Medical Imaging and Graphics*, 28(1-2):61–68, 2004.

[76] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv preprint arXiv:1906.02415*, 2019.

[77] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015.

[78] Prasanna K Sahoo, SAKC Soltani, and Andrew KC Wong. A survey of thresholding techniques. *Computer vision, graphics, and image processing*, 41(2):233–260, 1988.

[79] Gerald Schaefer, Maher I Rajab, M Emre Celebi, and Hitoshi Iyatomi. Colour and contrast enhancement for improved skin lesion segmentation. *Computerized Medical Imaging and Graphics*, 35(2):99–104, 2011.

[80] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *Toward category-level object recognition*, pages 127–144. Springer, 2006.

[81] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

[82] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[83] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[84] Scott E Umbaugh, Randy H Moss, and William V Stoecker. Automatic color segmentation of images with application to detection of variegated coloring in skin tumors. *IEEE Engineering in Medicine and Biology Magazine*, 8(4):43–50, 1989.

[85] Scott E Umbaugh, Randy H Moss, and William V Stoecker. An automatic color segmentation algorithm with application to identification of skin tumor borders. *Computerized Medical Imaging and Graphics*, 16 (3):227–235, 1992.

[86] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.

[87] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23 (7):903, 2004.

[88] Paul Wighton, Tim K Lee, Harvey Lui, David I McLean, and M Stella Atkins. Generalizing common tasks in automated skin lesion diagnosis. *IEEE Transactions on Information Technology inBiomedicine*, 15(4):622, 2011.

[89] Lang Xu, Marcel Jackowski, A Goshtasby, D Roseman, S Bines, C Yu, Akshaya Dhawan, and A Huntley. Segmentation of skin cancer images. *Image and Vision Computing*, 17(1):65–74, 1999.

[90] Y. Xue, T. Xu, and X. Huang. Adversarial learning with multi-scale loss for skin lesion segmentation. In *International Symposium on Biomedical Imaging*, pages 859–863, 2018.

[91] Xiaojing Yuan, Ning Situ, and George Zouridakis. A narrow band graph partitioning method for skin lesion segmentation. *Pattern Recognition*, 42(6):1017–1028, 2009.

[92] Yading Yuan. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. *arXiv preprint arXiv:1703.05165*, 2017.

[93] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[94] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. pages 1529–1537, 2015.

[95] Huiyu Zhou, Gerald Schaefer, M Emre Celebi, Faquan Lin, and Tangwei Liu. Gradient vector flow with mean shift for skin lesion segmentation. *Computerized Medical Imaging and Graphics*, 35(2):121–127, 2011.

[96] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

Appendices

Appendix A

Complete Results of Inter-Annotator Agreement Experiments

Table A.1 presents all the results discussed along with Chapter 4. The table includes the train and test datasets, the conditionings used during training and testing and the *mean* ± *std* Jaccard Index and Jaccard Index with 0.65 Threshold.

Results for Inter-Annotator Agreement Experiments

Dataset Train	Dataset Test	Cond. Train	Cond. Test	Jacc	Jacc Thr
ISIC Clean	Dermofit	Convex Hull	Convex Hull	0.717 ± 0.004	0.566 ± 0.010
ISIC Clean	Dermofit	Convex Hull	Opening	0.693 ± 0.009	0.504 ± 0.020
ISIC Clean	Dermofit	Convex Hull	No Cond.	0.684 ± 0.010	0.477 ± 0.012
ISIC Clean	Dermofit	Opening	Convex Hull	0.719 ± 0.008	0.571 ± 0.036
ISIC Clean	Dermofit	Opening	Opening	0.724 ± 0.004	0.578 ± 0.025
ISIC Clean	Dermofit	Opening	No Cond.	0.720 ± 0.004	0.565 ± 0.020
ISIC Clean	Dermofit	No Cond.	Convex Hull	0.713 ± 0.018	0.559 ± 0.054
ISIC Clean	Dermofit	No Cond.	Opening	0.713 ± 0.018	0.532 ± 0.073
ISIC Clean	Dermofit	No Cond.	No Cond.	0.708 ± 0.018	0.529 ± 0.072
ISIC Clean	ISIC Titans	Convex Hull	Convex Hull	0.762 ± 0.008	0.749 ± 0.029
ISIC Clean	ISIC Titans	Convex Hull	Opening	0.757 ± 0.010	0.727 ± 0.031
ISIC Clean	ISIC Titans	Convex Hull	No Cond.	0.746 ± 0.011	0.717 ± 0.031
ISIC Clean	ISIC Titans	Opening	Convex Hull	0.728 ± 0.016	0.618 ± 0.036
ISIC Clean	ISIC Titans	Opening	Opening	0.757 ± 0.017	0.726 ± 0.024
ISIC Clean	ISIC Titans	Opening	No Cond.	0.757 ± 0.016	0.743 ± 0.038
ISIC Clean	ISIC Titans	No Cond.	Convex Hull	0.722 ± 0.008	0.583 ± 0.007
ISIC Clean	ISIC Titans	No Cond.	Opening	0.748 ± 0.009	0.722 ± 0.032
ISIC Clean	ISIC Titans	No Cond.	No Cond.	0.748 ± 0.010	0.722 ± 0.032
ISIC Clean	PH2	Convex Hull	Convex Hull	0.852 ± 0.002	0.852 ± 0.002

Continuation of Results

Dataset Train	Dataset Test	Cond. Train	Cond. Test	Jacc	Jacc Thr
ISIC Clean	PH2	Convex Hull	Opening	0.808 ± 0.008	0.808 ± 0.008
ISIC Clean	PH2	Convex Hull	No Cond.	0.793 ± 0.009	0.793 ± 0.009
ISIC Clean	PH2	Opening	Convex Hull	0.845 ± 0.005	0.845 ± 0.005
ISIC Clean	PH2	Opening	Opening	0.836 ± 0.004	0.836 ± 0.004
ISIC Clean	PH2	Opening	No Cond.	0.826 ± 0.005	0.826 ± 0.005
ISIC Clean	PH2	No Cond.	Convex Hull	0.843 ± 0.003	0.843 ± 0.003
ISIC Clean	PH2	No Cond.	Opening	0.836 ± 0.007	0.836 ± 0.007
ISIC Clean	PH2	No Cond.	No Cond.	0.825 ± 0.008	0.825 ± 0.008
ISIC Full	Dermofit	Convex Hull	Convex Hull	0.700 ± 0.004	0.510 ± 0.014
ISIC Full	Dermofit	Convex Hull	Opening	0.678 ± 0.008	0.436 ± 0.038
ISIC Full	Dermofit	Convex Hull	No Cond.	0.669 ± 0.008	0.421 ± 0.035
ISIC Full	Dermofit	Opening	Convex Hull	0.695 ± 0.007	0.505 ± 0.038
ISIC Full	Dermofit	Opening	Opening	0.692 ± 0.010	0.482 ± 0.028
ISIC Full	Dermofit	Opening	No Cond.	0.687 ± 0.011	0.473 ± 0.026
ISIC Full	Dermofit	No Cond.	Convex Hull	0.698 ± 0.020	0.520 ± 0.061
ISIC Full	Dermofit	No Cond.	Opening	0.695 ± 0.020	0.506 ± 0.063
ISIC Full	Dermofit	No Cond.	No Cond.	0.690 ± 0.020	0.489 ± 0.058
ISIC Full	ISIC Titans	Convex Hull	Convex Hull	0.754 ± 0.014	0.723 ± 0.037
ISIC Full	ISIC Titans	Convex Hull	Opening	0.746 ± 0.014	0.730 ± 0.040
ISIC Full	ISIC Titans	Convex Hull	No Cond.	0.733 ± 0.014	0.718 ± 0.039
ISIC Full	ISIC Titans	Opening	Convex Hull	0.730 ± 0.005	0.623 ± 0.027
ISIC Full	ISIC Titans	Opening	Opening	0.753 ± 0.007	0.729 ± 0.031
ISIC Full	ISIC Titans	Opening	No Cond.	0.752 ± 0.008	0.715 ± 0.032
ISIC Full	ISIC Titans	No Cond.	Convex Hull	0.732 ± 0.009	0.622 ± 0.033
ISIC Full	ISIC Titans	No Cond.	Opening	0.754 ± 0.011	0.739 ± 0.036
ISIC Full	ISIC Titans	No Cond.	No Cond.	0.751 ± 0.011	0.736 ± 0.036
ISIC Full	PH2	Convex Hull	Convex Hull	0.855 ± 0.002	0.855 ± 0.002
ISIC Full	PH2	Convex Hull	Opening	0.810 ± 0.005	0.810 ± 0.005
ISIC Full	PH2	Convex Hull	No Cond.	0.794 ± 0.005	0.794 ± 0.005
ISIC Full	PH2	Opening	Convex Hull	0.851 ± 0.002	0.851 ± 0.002
ISIC Full	PH2	Opening	Opening	0.839 ± 0.003	0.839 ± 0.003
ISIC Full	PH2	Opening	No Cond.	0.828 ± 0.003	0.828 ± 0.003
ISIC Full	PH2	No Cond.	Convex Hull	0.850 ± 0.008	0.850 ± 0.008
ISIC Full	PH2	No Cond.	Opening	0.836 ± 0.006	0.836 ± 0.006
ISIC Full	PH2	No Cond.	No Cond.	0.825 ± 0.005	0.825 ± 0.005

Continuation of Results					
Dataset Train	Dataset Test	Cond. Train	Cond. Test	Jacc	Jacc Thr

Table A.1: Complete results for inter-annotator egreement experiments presented along with Chapter 4

Appendix B

Complete Results of Inter-Annotator Agreement Experiments

Table B.1 presents the results of our ANOVA for the experiments using the Inter-Annotator Agreement described along with Chapter 4. Note that all the main effects observed are strongly significant.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dataset_train	1	0.0018	0.0018	17.719	5.05e-05 ***
dataset_test	2	0.5094	0.25468	2511.244	<2e-16 ***
conditioning_train	2	0.0032	0.00162	16.004	7.18e-07 ***
conditioning_test	2	0.0044	0.00222	21.865	8.55e-09 ***
dataset_train:dataset_test	2	0.0045	0.00225	22.163	6.89e-09 ***
dataset_train:conditioning_train	2	0.0004	0.00019	1.897	0.1546
dataset_test:conditioning_train	4	0.0038	0.00096	9.461	1.16e-06 ***
dataset_train:conditioning_test	2	0.0001	0.00005	0.520	0.5956
dataset_test:conditioning_test	4	0.0107	0.00267	26.341	1.39e-15 ***
conditioning_train:conditioning_test	4	0.0091	0.00228	22.471	8.72e-14 ***
dataset_train:dataset_test:conditioning_train	4	0.0009	0.00023	2.312	0.0617 .
dataset_train:dataset_test:conditioning_test	4	0.0000	0	0.012	0.9997
dataset_train:conditioning_train:conditioning_test	4	0.0000	0.00001	0.086	0.9867
dataset_test:conditioning_train:conditioning_test	8	0.0003	0.00004	0.385	0.9269
dataset_train:dataset_test:conditioning_train:conditioning_test	8	0.0000	0	0.035	1.0000
Residuals	117	0.0119	0.0001		

Table B.1: ANOVA for the experiments presented in Chapter 4. We present all the evaluated experiments and the significance of our main results.