Diagnosing Skin Lesions from Images with Convolutional Neural Networks

by

Jeremy Glen Kawahara

M.Sc., Simon Fraser University, 2013
B.Sc., Vancouver Island University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY
Summer 2019

Approval

Name:	**Jeremy Glen Kawahara**
Degree:	**Doctor of Philosophy (Computing Science)**
Title:	**Diagnosing Skin Lesions from Images with Convolutional Neural Networks**
Examining Committee:	**Chair:** Steven Bergner University Research Associate
	Ghassan Hamarneh Senior Supervisor Professor
	Greg Mori Supervisor Professor
	Mark Drew Internal Examiner Professor Faculty of Applied Science
	M. Emre Celebi External Examiner Professor Department of Computer Science University of Central Arkansas
Date Defended:	**May 7th, 2019**

Abstract

Dermatological disorders are among the most common reasons for patients to visit general practitioners and are often diagnosed based on the visual properties of the affected skin. Machine classification of skin lesions from images offers the potential for a low cost, reproducible diagnosis that improves patients' access to dermatological expertise.

This thesis studies automated approaches to diagnose skin disorders based on visual cues within colour images. The thesis begins with a review of the existing literature on visual diagnosis of skin disorders, compares the reported performance of humans and machines, and discusses general limitations of image-based diagnosis. This thesis then proposes five novel approaches that rely on convolutional neural networks (CNNs) to diagnosis skin lesions.

The first two proposed works demonstrate that the parameters of a CNN, learned from non-skin images, transfer well to the tasks of skin lesion classification and image retrieval. The third work proposes a multi-resolution CNN architecture with end-to-end training that further increases classification accuracy. The final two works classify and localize several visual criteria that are commonly associated with melanoma from dermoscopy images, where a multi-task loss function in a multi-modal CNN architecture is proposed for classification, and a multi-label Dice score modified for imbalanced data is proposed to localize infrequently occurring melanoma-specific criteria. Finally, this thesis concludes with open questions that may benefit from further collaboration between dermatologists and computing researchers. This thesis demonstrates the potential role for CNNs as a common methodological building block to address the visual component of a variety of clinical problems within dermatology.

Keywords: skin lesions; convolutional neural networks; computer-aided diagnosis; image retrieval; detection

Acknowledgements

I am fortunate to have Prof. Ghassan Hamarneh as my senior supervisor. Thank you Prof. Hamarneh for being a strong example that I could learn from, for being a mentor, and for your constant encouragement and support. Thank you to my examination committee, Prof. Greg Mori, Prof. Mark Drew, Dr. Steven Bergner, and Prof. M. Emre Celebi, for all your time and feedback on this thesis.

Thank you to all my labmates at the Medical Image Analysis Lab. I learned so much from all our discussions. Special thanks to Dr. Colin J. Brown, Dr. Aïcha BenTaieb, and Kathleen P. Moriarty, for our brainstorming sessions and for your insights and feedback.

Thank you to my family and friends for your care. Finally, thank you to my parents for always supporting whatever I do.

Table of Contents

List of Tables

List of Figures

Nomenclature

Common Abbreviations

CNN Convolutional neural network

AUROC Area under the receiver operating characteristic curve

ROC Receiver operating characteristic

GPU Graphics processing unit

Dice-F$_1$ The Sørensen-Dice coefficient, also known as the F_1 score

Common Notation

ϕ A model, generally a neural network

θ The learned parameters of a model

x The input data to a model, typically an image

y The ground truth diagnosis for a lesion

$\hat{y}$ The predicted diagnosis for a lesion

z The ground truth dermoscopic criteria

$\hat{z}$ The predicted dermoscopic criteria

τ The ground truth seven-point criteria score

$\hat{\tau}$ The predicted seven-point criteria score

N The number of instances

$\mathcal{L}$ A loss function to minimize

$\mathbb{Z}$ Set of integers

Chapter 1

Introduction

1.1 Background and Motivations

Skin diseases are a common [142], widespread health concern [72], and diagnosing skin disorders can help determine appropriate treatment. Although histopathology analysis of a skin biopsy is considered to be the "gold standard" for identifying disorders such as melanoma [157], biopsy requires additional costs and can introduce further complications (e.g., infections [166]). A non-invasive diagnosis via patient consultation and visual inspection of the lesion can reduce costs and potential biopsy complications. As access to dermatologists may be limited in remote locations, *store-and-forward teledermatology*, where images of the lesion and patient history are sent to a remote physician, can potentially increase access to dermatological expertise [156]. Machine classification of dermatological disorders from images may further increase access to dermatological expertise through low cost, automated tools capable of diagnosing or screening skin disorders.

This thesis has both clinical and technical motivations. The *clinical* motivation of this thesis considers that with widespread internet usage and affordable high-quality mobile cameras, automated machine classification of skin images may offer a more accessible form of diagnosis and thus improve patients' treatments and outcomes. The *technical* motivation is to advance both state-of-the-art machine classification of skin conditions and those techniques that move towards interpretable diagnoses.

While many works focus on classifying dermoscopy images [15, 33] and identifying melanoma [99], the works in this thesis classify among several types of skin disorders using both clinical and dermoscopy images in order to approach a more realistic clinical setting. Methodologically, the convolutional neural network (CNN) [102] serves as a core common component that we build on in each work. This methodological choice is motivated by the success of CNNs when classifying the natural images of ImageNet [140], which share similar properties to dermatological images. In particular, given that there are thousands of known skin disease types [22], building specialized manually engineered features or pipelines to target each type of dermatological disorder will require significant human engineering effort.

The following two quotes from Bengio and LeCun [20], while referring to machine learning in general, are also applicable to dermatology:

> "Consider for example the problem of object recognition in computer vision: we could be interested in building recognizers for at least several thousand categories of objects. Should we have specialized algorithms for each?"
>
> "If there exist more general-purpose learning models, at least general enough to handle most of the tasks that animals and humans can handle, then searching for them may save us a considerable amount of labor in the long run."

The second common methodological choice we made was to repurpose the parameters of a CNN trained over natural images for dermatological images. This *transfer-learning*, or *pre-training*, approach addresses the lack of available labelled dermatological images for training the millions of parameters commonly found within CNNs, and often reduces the computational requirements to train a model.

1.2 Thesis Contributions

The following sections outline the chapters within the thesis. Each section briefly describes the chapter, highlights key contributions, and concludes with details about the resulting publication that corresponds to each work. The full details of each work are provided in the respective chapters.

1.2.1 Visual Diagnosis of Dermatological Disorders: Human and Machine Performance

Dermatologists have extensively researched how to diagnose conditions from a patient's history and the lesion's visual appearance. Computer vision researchers are attempting to encode this diagnostic ability into machines, and several recent studies now report machine level performance comparable with dermatologists.

Chapter 2 reviews machine approaches to classify skin images and considers their performance when compared to human dermatologists. Following an overview of common image modalities, dermatologists' diagnostic approaches, common tasks, and publicly available datasets, we discuss approaches to machine skin lesion classification. We then review works that directly compare human and machine performance. Finally, this chapter addresses the limitations and sources of errors in image-based skin disease diagnosis, applicable to both machines and dermatologists in a teledermatology setting.

Contribution

- Surveys the performance of human and machine skin lesion classification across multiple published dermatological works

1.2.2 Deep Features to Classify Skin Lesions

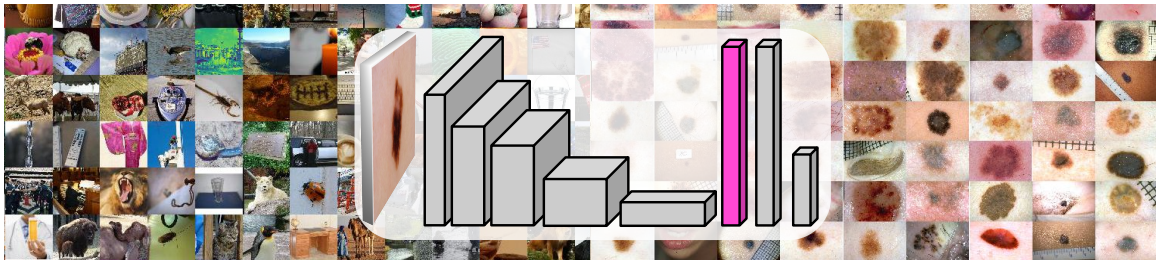


Figure 1.1: Chapter 3 demonstrates that the learned parameters of a CNN trained to classify a large dataset of non-skin images (*left side*) will generalize to skin lesion classification.

The number of labelled dermatology images is less than the number of labelled natural images (i.e., non-skin specific images), which makes applying supervised machine learning approaches challenging in dermatology. While natural images differ in appearance from dermatology images (Fig. 1.1), a CNN trained on one domain may generalize to another domain. Chapter 3 evaluates whether the parameters of a CNN trained on natural images generalizes well to dermatological images. To this end, we demonstrate that a linear classifier, trained on features extracted from a CNN pretrained on natural images, distinguishes among ten skin lesions with higher accuracy than previously published state-of-the-art results on the same dataset. Further, in contrast to competing works, our approach requires neither lesion segmentations nor complex preprocessing.

We gain additional improvements to accuracy using a per image normalization, a fully-convolutional neural network to extract multi-scale features, and by pooling over an augmented feature space. Compared to state-of-the-art, our proposed approach achieves a favourable accuracy of 85.8% over 5-classes (vs. 75.1%) with noticeable improvements in accuracy for underrepresented classes (e.g., 60% compared to 15.6%). Over the entire 10-class dataset of 1,300 images captured from a standard (non-dermoscopy) camera, our method achieves an accuracy of 81.8% outperforming the 67% accuracy previously reported.

Contributions

- First work to use the parameters from a CNN trained on natural images to perform multi-class (greater than two) classification of clinical dermatological images
- Outperforms the reported performance of using traditional engineered features

This chapter was accepted as a podium presentation at the 2016 International Symposium on Biomedical Imaging conference, where it was awarded runner-up for best student paper. To date, this work has garnered 75 citations.

> [85] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in Proceedings of the IEEE International Symposium on Biomedical Imaging, 2016, pp. 1397-1400. `https://doi.org/10.1109/ISBI.2016.7493528`

1.2.3 Graph Geodesics to Find Progressively Similar Skin Lesion Images

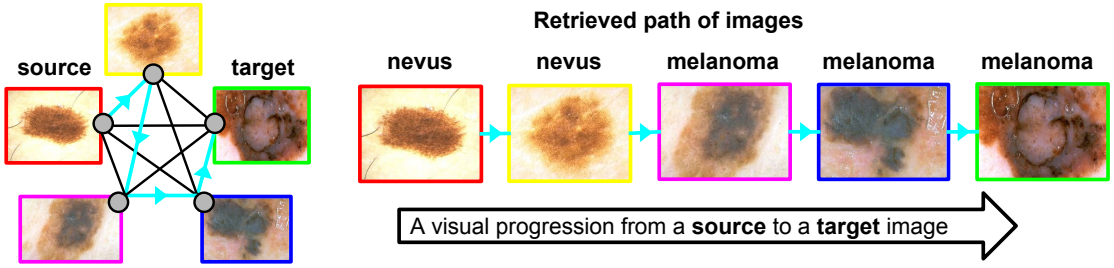


Figure 1.2: Chapter 4 presents a visualization approach that returns a path of visually similar images between a given source and target image.

As the previous chapter demonstrated the utility of pretrained parameters to classify skin lesions, Chapter 4 relies on pretrained parameters to measure the similarity among images. Rather than traditional image retrieval, which retrieves images similar to a single query image, we propose a graph-based approach to visualize a progression of similar skin images between *pairs* of images. This visualization approach may offer new ways to explore and gain insights into datasets of skin images (e.g., find reference images that are visually challenging to classify, show how a benign lesion may visually progress to a malignancy).

In our graph, a node represents both a clinical and dermoscopy image of the same lesion, and an edge between nodes captures the visual dissimilarity between lesions (Fig. 1.2). Dissimilarity is computed by comparing the image responses of a pretrained CNN. We compute the geodesic/shortest path between nodes to determine a path of progressively visually similar skin lesions. To quantitatively evaluate the quality of the returned path, we propose metrics to measure the number of transitions with respect to the lesion diagnosis, and the progression with respect to the clinical 7-point checklist. Compared to baseline experiments, our approach makes improvements to the quality of the returned paths.

Contributions

- First work to apply geodesic paths to visualize skin lesion images
- Proposed metrics to quantify the quality of retrieved paths

This chapter was published in the MICCAI Workshop on Graphs in Biomedical Image Analysis and was accepted as an oral presentation.

[92] J. Kawahara, K. P. Moriarty, and G. Hamarneh, "Graph geodesics to find progressively similar skin lesion imagess," in International Workshop on Graphs in Biomedical Image Analysis, 2017, vol. 10551, pp. 31-41. https://doi.org/10.1007/978-3-319-67675-3_4

1.2.4 Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-lesion Trained Layers

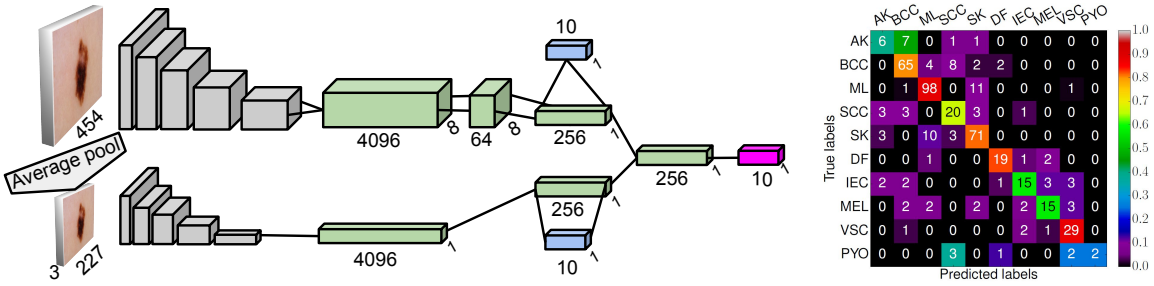


Figure 1.3: Chapter 5 proposes a CNN that is trained on multiple resolutions of the same image simultaneously and predicts 10 types of skin diseases (*purple block*).

The previous two chapters demonstrated that pretrained parameters transfer well when classifying and retrieving skin images; however, these pretrained CNN's parameters did not learn from the targeted skin domain. In Chapter 5, we fine-tune the pretrained parameters of a CNN to classify skin lesions using our proposed CNN architecture. While traditional CNNs are generally trained on a single resolution image, our CNN is composed of multiple tracts, where each tract analyzes the image at a different resolution *simultaneously* and learns interactions across multiple image resolutions using the *same* field-of-view. We convert a CNN, pretrained on a single resolution (gray blocks in Fig. 1.3), to work for multi-resolution input (green blocks in Fig. 1.3, details provided in the chapter). The entire network is fine-tuned in a fully learned end-to-end optimization with auxiliary loss functions (blue blocks in Fig. 1.3). We show how our proposed novel multi-tract network yields higher classification accuracy, outperforming state-of-the-art multi-scale approaches when compared over a public skin lesion dataset.

Contribution

- First end-to-end training of a pretrained multi-resolution CNN for dermatology images

This chapter was published in the MICCAI Workshop on Machine Learning In Medical Imaging, was accepted as an oral presentation, and has 31 citations to date.

[90] J. Kawahara and G. Hamarneh, "Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers," in International Workshop on Machine Learning In Medical Imaging, 2016, pp. 164-171. `https://doi.org/10.1007/978-3-319-47157-0_20`

1.2.5 Seven-Point Checklist and Skin Lesion Classification using Multi-task Multimodal Neural Nets

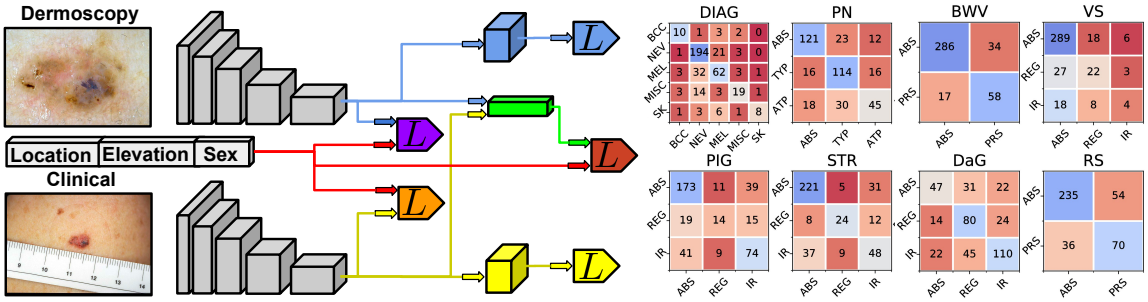


Figure 1.4: Chapter 6 proposes to diagnose a skin lesion and classify seven dermoscopic criteria using a dermoscopy image, a clinical image, and patient meta-data. The CNN is robust to missing data at inference time due to multiple loss functions (*blocks with L*).

Detecting visual characteristics of a lesion commonly associated with a disease is another approach to diagnosis, often used when considering melanoma in dermoscopy. Chapter 6 proposes a multi-task deep CNN, trained on multi-modal data (clinical and dermoscopy images, and patient meta-data), to classify the 7-point melanoma checklist criteria and perform skin lesion diagnosis (Fig. 1.4). Our neural network is trained using several multi-task loss functions, where each loss considers different combinations of the input modalities (e.g., the blue L-block in Fig. 1.4 is a loss that is only a function of dermoscopy images), which allows our model to be robust to missing data at inference time. Our final model classifies both the 7-point checklist and skin condition diagnosis, produces multi-modal feature vectors suitable for image retrieval, and localizes clinically discriminative regions. We benchmark our approach using 1,011 lesion cases and report comprehensive results over the 7-point criteria and diagnosis.

Contributions

- Single model to diagnose diseases and classify all the 7-point checklist types
- The multi-modal model is robust to missing data at inference time
- Dataset (images, metadata, and labels) is publicly available with over 70 data requests to date: `http://derm.cs.sfu.ca`

This chapter was published in the IEEE Journal of Biomedical and Health Informatics:

> [88] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538-546, 2019. `https://doi.org/10.1109/JBHI.2018.2824327`

1.2.6 Fully Convolutional Neural Networks to Detect Dermoscopic Criteria

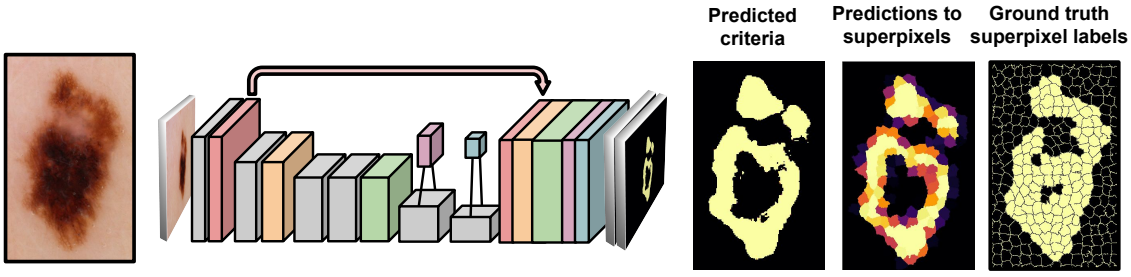


Figure 1.5: Chapter 7 describes a CNN that detects dermoscopic criteria.

The previous chapter described an approach to *classify* dermoscopic criteria associated with melanoma; however, *detection*, which both localizes and identifies regions of the image that contain dermoscopic criteria, may lead to more interpretable diagnoses. Chapter 7 reformulates the task of classifying dermoscopic criteria within superpixels as a segmentation problem and proposes a fully-convolutional neural network to detect dermoscopic criteria from dermoscopy images (Fig. 1.5). Our neural network architecture uses interpolated feature maps from several intermediate network layers (red arrow in Fig. 1.5) and addresses imbalanced labels by minimizing a negative multi-label Dice-F_1 score, where the score is computed across the mini-batch for each label.

Our approach ranked first place in the *2017 ISIC-ISBI Part 2: Dermoscopic Feature Classification Task* challenge over both the provided validation and test datasets, achieving a 89.5% area under the receiver operator characteristic curve score. We show how simple baseline models can outrank state-of-the-art approaches when using the official metrics of the challenge, and propose to use a fuzzy Jaccard Index that ignores the empty set (i.e., masks devoid of positive pixels) when ranking models. Our results suggest that the classification of dermoscopic criteria can be effectively approached as a segmentation problem, and the current metrics used to rank models may not well capture the efficacy of the model. We plan to make our trained model and code publicly available.

Contributions

- Modified Dice loss function to addresses multiple types of class-imbalance
- Approach ranked first on a public competition

This chapter was published in the IEEE Journal of Biomedical and Health Informatics:

> [91] J. Kawahara and G. Hamarneh, "Fully convolutional neural networks to detect clinical dermoscopic features," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 578-585, 2019. `https://doi.org/10.1109/JBHI.2018.2831680`

1.3 Works Excluded From Thesis

The following is a list of peer-reviewed publications I was involved with over the course of my PhD studies but were not selected to be a part of the thesis.

[93] **J. Kawahara**, J. M. Peyrat, J. Abinahed, O. Al-Alao, A. Al-Ansari, R. Abugharbieh, and G. Hamarneh, "Automatic labelling of tumourous frames in free-hand laparoscopic ultrasound video," in Medical Image Computing and Computer Assisted Intervention, 2014, vol. 8674 LNCS, no. Part II, pp. 676-683. https://doi.org/10.1007/978-3-319-10470-6_84

[69] G. Hamarneh, A. Amir-Khalili, M. S. Nosrati, I. Figueroa, **J. Kawahara**, et al.,"Towards multi-modal image-guided tumour identification in robot-assisted partial nephrectomy," in IEEE Middle East Conference on Biomedical Engineering, 2014, pp. 159-162. https://doi.org/10.1109/MECBME.2014.6783230

[89] **J. Kawahara** and G. Hamarneh, "Image content-based navigation of skin conditions," in World Congress of Dermatology, 2015, [abstract]. https://www.cs.sfu.ca/~hamarneh/ecopy/wcd2015a.pdf

[21] A. BenTaieb, **J. Kawahara**, and G. Hamarneh, "Multi-loss convolutional networks for gland analysis in microscopy," in IEEE International Symposium on Biomedical Imaging, 2016, pp. 642-645. https://doi.org/10.1109/ISBI.2016.7493349

[154] S. A. Taghanaki, **J. Kawahara**, B. Miles, and G. Hamarneh, "Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification," Computer Methods and Programs in Biomedicine, vol. 145, pp. 85-93, 2017. http://dx.doi.org/10.1016/j.cmpb.2017.04.012

[86] **J. Kawahara***, C. J. Brown*, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh, "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment," NeuroImage, vol. 146, no. Feb, pp. 1038-1049, 2017. https://doi.org/10.1016/j.neuroimage.2016.09.046 **Joint first authors*
https://github.com/jeremykawahara/ann4brains

[124] Z. Mirikharaji, S. Izadi, **J. Kawahara**, and G. Hamarneh, "Deep auto-context fully convolutional neural network for skin lesion segmentation," in IEEE International Symposium on Biomedical Imaging, 2018, pp. 877-880. https://doi.org/10.1109/ISBI.2018.8363711

[79] S. Izadi, Z. Mirikharaji, **J. Kawahara**, and G. Hamarneh, "Generative adversarial networks to segment skin lesions," in IEEE International Symposium on Biomedical Imaging, 2018, pp. 881-884. https://doi.org/10.1109/ISBI.2018.8363712

[28] C. J. Brown*, **J. Kawahara***, and G. Hamarneh, "Connectome priors in deep neural networks to predict autism," in IEEE International Symposium on Biomedical Imaging, 2018, pp. 110-113. https://doi.org/10.1109/ISBI.2018.8363534 **Joint first authors*

[172] Y. Yan, **J. Kawahara**, and G. Hamarneh, "Melanoma recognition via visual attention," in Information Processing in Medical Imaging, 2019, vol. 11492 LNCS, pp. 793-804. https://doi.org/10.1007/978-3-030-20351-1_62

Table 1.1: For each chapter, the clinical tasks, corresponding publication (*Pub.*), and a description of the dataset used is shown. The *Derm.* and *Clinic* columns indicate dermoscopy and clinical images, respectively. N_{imgs} indicates the number of available images, and N_{class} indicates the number of skin lesion categories.

Ch.	Tasks	Pub.	Dataset	Derm.	Clinic	N_{imgs}	N_{class}
3	Classify disease	[85]	Dermofit [12]		✓	1,300	10
4	Retrieve images	[92]	Atlas of Derm. [88]	✓	✓	2,018	15
5	Classify disease	[90]	Dermofit [12]		✓	1,300	10
6	Classify disease	[88]	Atlas of Derm. [88]	✓	✓	2,018	5
	Classify 7 criteria						19
	Retrieve images						24
7	Detect criteria	[91]	ISIC-2017 [41]	✓		2,000	4

1.4 Summary of Contributions

Overall, this thesis proposes approaches to advance machine diagnosis of skin lesions from images, where we use the pretrained CNN as a common building block. The pretrained CNN serves to initialize a model with parameters already well suited for classifying natural images and can help reduce the over-fitting problem that can occur when training a CNN on a relatively small dataset. The specific pretrained CNN used differs across chapters as new pretrained models capable of extracting more discriminative features became available. We modify the architecture of the traditional CNN to suit a variety of clinical tasks (e.g., diagnose diseases, detect visual properties of melanoma), with experiments spanning across three distinct dermatological datasets. In addition, we make both code [91] and data [88] available, and have competed in public skin lesion segmentation and clinical feature detection competitions, placing fourth and first place, respectively.

1.5 Thesis Outline

Chapter 2 presents a general overview of common types of dermatological images, datasets, and tasks. We survey typical approaches that human dermatologist and machines use to diagnose skin conditions, compare the performance across multiple human and machine dermatological studies, and discuss limitations of image-based diagnosis systems. Chapter 3 through Chapter 7 presents the proposed approaches to automated skin lesion diagnosis, as outlined in Section 1.2, where Table 1.1 summarizes the tasks and datasets used in each chapter. Finally, Chapter 8 summarizes the proposed works and contributions and suggests future directions.

Chapter 2

Visual Diagnosis of Dermatological Disorders: Human and Machine Performance

2.1 Introduction

Skin disorders are the most frequent reason to visit a general practitioner in studied populations [142] and are a recognized global health burden [72]. In 2013, approximately one in four Americans saw a physician for at least one skin condition [108]. As correctly diagnosing, or classifying, skin conditions can help narrow treatment options, dermatologists have extensively researched how to classify skin conditions from a patient's history and the visual properties of skin lesions. However, skin diseases are difficult to diagnose [143], and studies suggest an unmet demand for dermatologists [94]. To alleviate these challenges, computer vision researchers are attempting to encode this diagnostic ability into machines [99], which may lead to more reproducible and accessible diagnoses in under-served communities.

The following section provides an overview of the common imaging modalities, tasks, typical diagnostic approaches used by dermatologists, and common datasets and metrics used to evaluate the performance of automated skin disease diagnosis. Section 2.2 reviews trends in machine approaches to classify skin diseases. Section 2.3 presents works that directly compare humans and machine skin disease classification. Finally, Section 2.4 discusses the performance of humans and machines.

2.1.1 Non-Invasive Imaging Modalities of the Skin

The two common non-invasive imaging modalities to acquire skin images are clinical and dermoscopy images (Fig. 2.1). *Clinical* images capture what is seen with the unaided human eye and can be acquired at varying fields-of-view using non-standard cameras. *Dermoscopy* (also referred to as epiluminescence microscopy [43]) images show a magnified view of intra- and sub-epidermal structures and are acquired using a dermatoscope (Fig. 2.2a), which offers

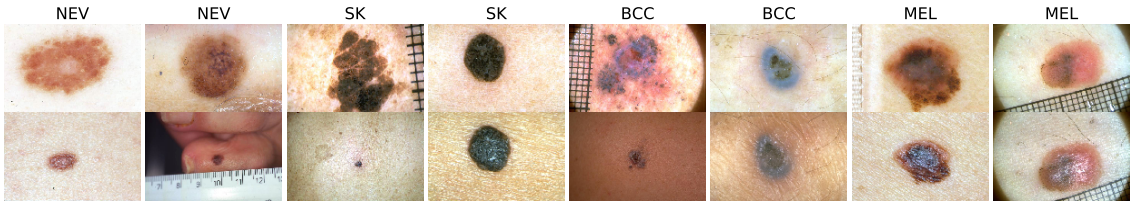


Figure 2.1: The same lesion (*column*) can be captured as a dermoscopy (*top row*) and a clinical image (*bottom row*). Dermoscopy offers a more standardized acquisition, while clinical images can capture a wider field of view. These sampled images [5] show the variability of some common lesions, where nevi (NEV) and seborrheic keratoses (SK) are benign conditions, and basal cell carcinoma (BCC) and melanoma (MEL) are common cancers.

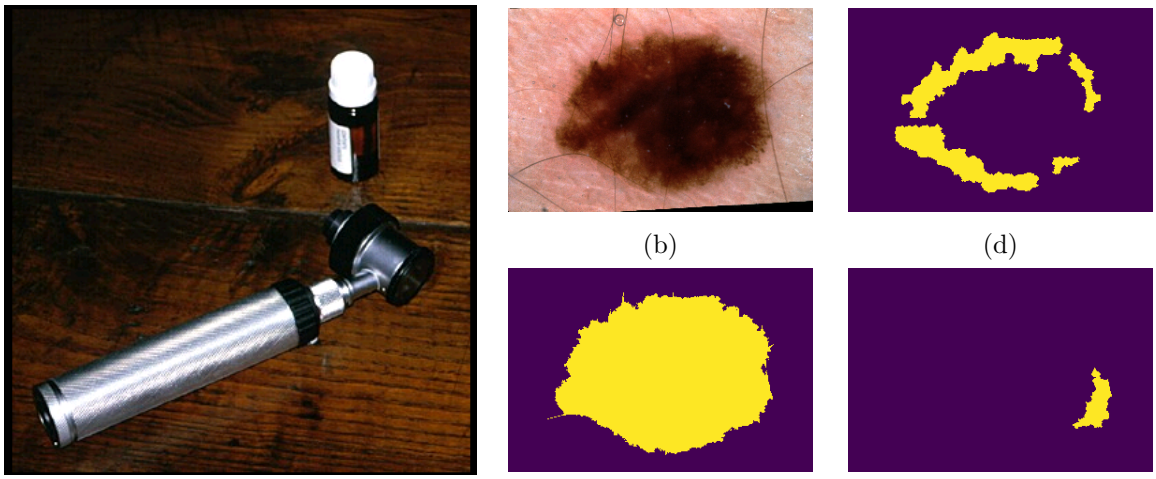
a more controlled field-of-view. Dermoscopy images are commonly used to help differentiate benign from malignant lesions, whereas clinical images, with their flexible field-of-view, are more commonly used to image general skin diseases. While other non-invasive imaging modalities, such as ultrasound, have been used for skin lesion diagnosis [168], this survey focuses on clinical and dermoscopy images. For this report, non-image information acquired from the patient is defined as the *patient history*, which includes factors such as patient age, sex, lesion location, family history, and environmental factors.

2.1.2 Diagnosing Skin Diseases

Diagnosing skin diseases is complicated. There are at least 3,000 identified varieties of skin diseases [22] with a prevalence that varies by condition. The "gold standard" for skin disease diagnosis is determined through a biopsy, where a portion of the affected skin specimen is extracted and sent to dermatopathologists for analysis [143]. However, biopsy requires additional time and cost to extract and analyze the lesion, and may introduce potential complications to the patient. Wahie et al. [166] reported that 29% of patients had complications after a skin biopsy, mainly as a result of infection. Thus dermatologists may avoid biopsy in cases with well-recognized symptoms and instead rely on data collected non-invasively.

Dermatologists consider a variety of factors in their diagnoses, including patient history and the appearance (e.g., morphology, colour, textures) of the affected skin region. Entire textbooks describe approaches to diagnose skin lesions (e.g., [10]), where the methods are often specific to distinct types of dermatological conditions. For example, to classify skin diseases that manifest as stains on the skin, flowcharts that encode visual properties, patient history, and the lesion's location on the body can aid in the diagnosis [45].

Melanoma, which accounted for 41% of skin related deaths in the United States in 2013 [108], receives special attention due to the mortality risk. To aid less experienced clinicians in recognizing melanoma from benign lesions, rule-based diagnostic systems have been proposed, such as the ABCD rule [129] and the 7-point checklist [9]. These simplified rule-



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

Figure 2.2: (a) An image of a dermatoscope [5]. (b-e) Common dermatological tasks: (b) *classify* the observed skin lesion directly from the image; (c) *segment* the lesion from the background; and, *detect* the presence of dermoscopic criteria (e.g., (d) pigment network and (e) streaks) within the image.

based systems produce a melanoma score based on the physician recognizing the presence of melanoma-specific morphological characteristics within the lesion.

General practitioners receive less training in dermatology than dermatologists and are often the first point of contact for skin conditions. Rübsam et al. [139] found that general practitioners reported diagnosing dermatological problems using various strategies: visual recognition, testing of different treatments, and refining the diagnosis via asking additional questions. Sellheyer et al. [143] reported that dermatologists correctly diagnosed roughly twice the number of cases when compared to non-dermatologists, using the histopathology diagnoses as the reference.

2.1.3 Common Non-Invasive Dermatology Tasks

Classify Skin Diseases

As previously discussed, to diagnose or classify a skin disease, a physician or machine predicts the type of skin disease by analyzing the patient's history, the visual properties of skin lesions, or both. Automated machine approaches to classify skin diseases is discussed in detail in Section 2.2. While it is possible to estimate the disease class from the image directly, this process, especially during machine classification, has traditionally been broken into the following sub-tasks.

Classify Dermoscopic Criteria

The existence of certain visual properties within a lesion may indicate a condition. For example, the presence of certain dermoscopic criteria (such as an atypical pigment network or irregular streaks) within a lesion is indicative of melanoma. Thus, one approach to classify melanoma is to classify dermoscopic criteria known to be associated with melanoma. If a lesion contains a number of these criteria, a diagnosis of melanoma can be inferred [9]. Approaches to classify dermoscopic criteria are discussed in Section 2.2.2.

Lesion Segmentation

Lesion segmentation, i.e., delineating the boundary of a lesion in an image (Fig. 2.2c), allows for lesion properties to be carefully measured, and is often used to extract image features that rely on knowing the border of the lesion. Several of the works discussed in Section 2.2 segment the lesion prior to classification.

Detect Dermoscopic Criteria

A specific dermoscopic criteria (e.g., streaks, which are associated with melanoma) can be both localized and classified (Fig. 2.2d and Fig. 2.2e). While this task is similar to classifying dermoscopic criteria (Sec. 2.1.3), detecting dermoscopic criteria requires localization. This task may allow physicians to localize those areas containing disease-specific criteria.

Artefact Removal

Artefact removal involves discarding potentially confounding properties from the images, and is often a preprocessing step that precedes the aforementioned tasks. For example, applying colour constancy to control for illumination [16], and removing specular highlights [114] or hair [2, 126] from images may improve lesion segmentation or classification.

2.1.4 Common Skin Condition Image Datasets

As diagnostic difficulty varies by image and type of condition, standardized datasets provide a valuable way to benchmark different approaches. Here we discuss commonly used and publicly available datasets suitable for classifying skin conditions from images.

Atlas of Dermoscopy

The Atlas of Dermoscopy, also know as the EDRA atlas, was originally released as a tool to instruct physicians to diagnose skin lesions and recognize dermoscopic criteria related to melanoma [5]. This dataset provides 1,011 cases of skin lesions, with corresponding clinical and dermoscopy images for nearly every case, patient history (e.g., age, sex), and ground truth diagnoses and dermoscopic criteria labels. Table 2.1 provides details on the number

of cases available for each dermoscopic criteria and diagnosis. We have made this dataset available online [87] and further dataset details are provided in Chapter 6.

Table 2.1: Details of the Atlas of Dermoscopy dataset. The two left columns show the labels for each criteria in the 7-point checklist. The right column shows the labels that correspond to the overall diagnoses. The *7pt* column indicates the contribution to the 7-point checklist score, where a non-zero score indicates a criteria label associated with melanoma. The *#cases* column indicates the number of cases with the specific label.

Name	7pt	#cases	Name	7pt	#cases	Name	#cases
1. *Pigment Network*			5. *Vascular Structures*			*Diagnosis*	
absent	0	400	absent	0	823	Basal Cell Carcinoma	42
typical	0	381	arborizing	0	31	Blue Nevus	28
atypical	2	230	comma	0	23	Clark Nevus	399
2. *Regression Structures*			hairpin	0	15	Combined Nevus	13
absent	0	758	within regression	0	46	Congenital Nevus	17
blue areas	1	116	wreath	0	2	Dermal Nevus	33
white areas	1	38	dotted	2	53	Recurrent nevus	6
combinations	1	99	linear irregular	2	18	Reed or Spitz Nevus	79
3. *Pigmentation*			6. *Dots and Globules*			Melanoma	252
absent	0	588	absent	0	229	Dermatofibroma	20
diffuse regular	0	115	regular	0	334	Lentigo	24
localized regular	0	3	irregular	1	448	Melanosis	16
diffuse irregular	1	265	7. *Streaks*			Miscellaneous	8
localized irregular	1	40	absent	0	653	Vascular Lesion	29
4. *Blue Whitish Veil*			regular	0	107	Seborrheic Keratosis	45
absent	0	816	irregular	1	251		
present	2	195				*Total Cases*	*1011*

Dermofit Image Library

The Dermofit Image Library [12] is available online [155] and consists of 1,300 clinical images covering 10 classes of skin lesions (described in Table 2.2). Images are captured in a standardized way, controlling for illumination and distance to the lesion. Manually segmented lesions are also available.

Table 2.2: The number of images for each skin disease type in the Dermofit Image Library.

Name	# imgs	Name	# imgs
Actinic Keratosis	45	Malignant Melanoma	76
Basal Cell Carcinoma	239	Melanocytic Nevus	331
Dermatofibroma	65	Pyogenic Granuloma	24
Haemangioma	97	Seborrhoeic Keratosis	257
Intraepithelial Carcinoma	78	Squamous Cell Carcinoma	88

Table 2.3: The diagnosis, dermoscopic criteria, and the number of images with each label in the PH^2 dataset.

Name	# imgs	Name	# imgs	Name	# imgs
Diagnosis		3. *Dots/Globules*		6. *Asymmetry*	
Common Nevus	80	Absent	87	Fully Symmetric	117
Atypical Nevus	80	Atypical	59	Asymmetry in One Axis	31
Melanoma	40	Typical	54	Fully Asymmetry	52
1. *Pigment Network*		4. *Streaks*		7. *Colors*	
Atypical	116	Absent	170	White	19
Typical	84	Present	30	Red	10
2. *Blue Whitish Veil*		5. *Regression Areas*		Light-Brown	139
Absent	164	Absent	175	Dark-Brown	156
Present	36	Present	25	Blue-Gray	38
				Black	42

PH^2

PH^2 is a publicly available [57] dataset of 200 dermoscopy images of skin lesion. Each lesion was manually segmented and expertly labeled with a diagnosis and seven dermoscopic criteria [120, 121]. These dermoscopic criteria are a subset of the 7-point checklist [9] and includes additional criteria relevant to other diagnostic procedures (e.g., ABCD rule [129]). Table 2.3 shows the number of images labeled with each diagnosis and dermoscopic criteria.

ISIC Challenge

The ISIC Challenge is a public dermatology competition with three tasks: segment lesions; detect dermoscopic criteria; and classify lesions (as described in Section 2.1.3). The challenge has run in 2016 [66], 2017 [41], and 2018 [39]. The dataset contains dermoscopy images, lesion segmentation masks, and dermoscopic criteria masks. Each task has standard evaluation metrics and training, validation, and testing dataset partitions. In 2018, the ISIC skin lesion classification challenge used the HAM10000 dataset [162] as the training set, which provides 10,015 dermoscopy images covering seven diagnosis categories. Table 2.4 shows the number of images with each diagnosis label in the HAM10000 dataset.

Table 2.4: The diagnosis labels for the HAM10000 dataset.

Name	# imgs
Actinic Keratosis & Intraepithelial Carcinoma	327
Basal Cell Carcinoma	514
Benign Keratosis	1099
Dermatofibroma	115
Melanoma	1113
Melanocytic Nevus	6705
Vascular Lesion	142

SD Clinical Dataset

The SD-198 dataset [173] consists of 6,584 clinical images covering 198 fine-grained categories of skin diseases, where each category has between 10 and 60 images [151]. Images were collected from the website DermQuest [48]. SD-128 is a subset of SD-198 and consists of 5,619 clinical images, where each class has at least 20 images.

Others

The *Skin Cancer Detection* dataset [164] provides 119 melanoma and 87 non-melanoma clinical images along with lesion segmentation masks, which were gathered from two online sources: the Dermatology Information System [49] and DermQuest [48]. The *MED-NODE* dataset [47, 65] provides 70 melanoma and 100 nevi clinical images. The *Melanoma Classification Benchmark* [25, 26] sourced 100 dermoscopy images from the ISIC challenge [66] and 100 clinical images from the MED-NODE dataset [47] such that for each type of image, 80 images are of benign nevi and 20 are of melanoma. The diagnostic performance of dermatologists over the same dataset is also provided (157 dermatologists for dermoscopy images, 145 dermatologists for clinical images).

2.1.5 Common Metrics for Classification

Measuring the performance of a classifier on a diagnostic task is complicated as there are multiple classes of diseases, datasets are often imbalanced, and the clinical penalty for misdiagnosis may differ.

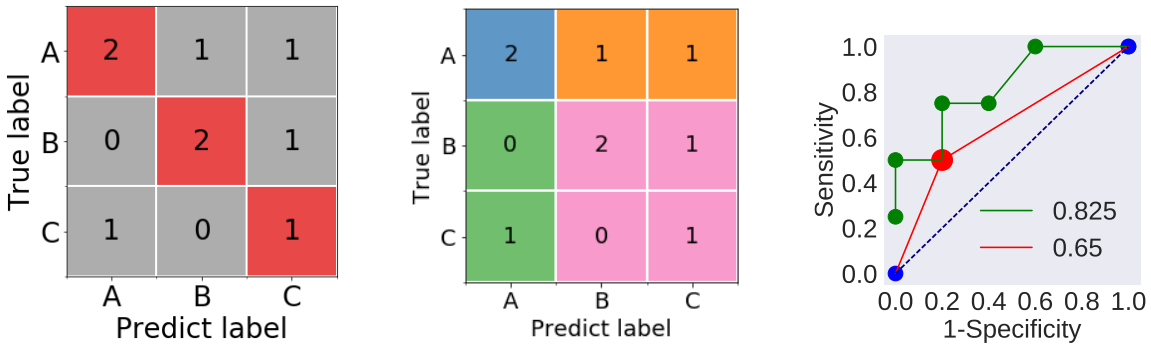
Accuracy is a common metric that measures the fraction of cases where the predicted diagnoses $\hat{y}$ correctly matches the true diagnoses y. Some clinical studies (e.g., [169]) report results that include a differential diagnosis (i.e., when a physician makes more than one disease diagnosis), where the prediction is considered correct if any of the K diagnoses match the true diagnosis. The top-K accuracy is defined as,

$$\text{accuracy}(\hat{y}, y, K) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \delta(\hat{y}^{(i)}, y_k^{(i)}) \tag{2.1}$$

where there are N cases; $\delta(a, b)$ is the Kronecker delta function which returns 1 if $a = b$, else 0; $y^{(i)}$ is the true diagnosis for the i-th case; and, $\hat{y}_k^{(i)}$ is the k-th predicted differential diagnosis of the i-th case. Given a confusion matrix of the predicted and true labels, the top-1 accuracy can be computed by dividing the sum of the diagonal values by N (Fig. 2.3a).

Other common metrics for classification problems are sensitivity,

$$\text{sensitivity}(\hat{y}, y, c) = \frac{\text{TP}(\hat{y}, y, c)}{\text{TP}(\hat{y}, y, c) + \text{FN}(\hat{y}, y, c)} \tag{2.2}$$



(a) (b) (c)

Figure 2.3: Computing common classification metrics. (a) Given a confusion matrix of three-classes, accuracy is computed by dividing the sum of the diagonal cells (red) with the total number of samples within all cells. (b) Considering "A" as a positive class, the blue cell indicates true positives, orange cells indicate false negatives, green cells indicate false positives, and pink cells indicate true negatives. (c) The AUROC curve with respect to a single class "A". Discrete predictions (e.g., red point) may have a different ROC curve than probabilistic predictions (*green line*). The values in the legend are the AUROC of the corresponding curves.

and specificity,

$$\text{specificity}(\hat{y}, y, c) = \frac{\text{TN}(\hat{y}, y, c)}{\text{TN}(\hat{y}, y, c) + \text{FP}(\hat{y}, y, c)} \tag{2.3}$$

where each metric is computed with respect to a class label c (e.g., a diagnosis y may have C possible class labels). Given c as the positive class label, the number of true positives, false positives, true negatives, and false negatives are computed as,

$$\text{TP}(\hat{y}, y, c) = \sum_{i=1}^{N} \Big(\delta(\hat{y}^{(i)}, c) \cdot \delta(y^{(i)}, c) \Big) \tag{2.4}$$

$$\text{FP}(\hat{y}, y, c) = \sum_{i=1}^{N} \Big(\delta(\hat{y}^{(i)}, c) \cdot (1 - \delta(y^{(i)}, c)) \Big) \tag{2.5}$$

$$\text{TN}(\hat{y}, y, c) = \sum_{i=1}^{N} \Big((1 - \delta(\hat{y}^{(i)}, c)) \cdot (1 - \delta(y^{(i)}, c)) \Big) \tag{2.6}$$

$$\text{FN}(\hat{y}, y, c) = \sum_{i=1}^{N} \Big((1 - \delta(\hat{y}^{(i)}, c)) \cdot \delta(y^{(i)}, c) \Big) \tag{2.7}$$

respectively. Fig. 2.3b shows an example using a confusion matrix.

Another metric used to measure the performance over a public skin dataset is balanced accuracy. This metric is equivalent to the sensitivity averaged across classes,

$$\overline{\text{sensitivity}}(\hat{y}, y) = \frac{1}{C} \sum_{c=1}^{C} \text{sensitivity}(\hat{y}^{(i)}, y^{(i)}, c) \tag{2.8}$$

where C is the number of unique classes. The averaged sensitivity assumes an equal importance for each class and may be more suitable for imbalanced datasets than accuracy (Eq. 2.1), as accuracy assumes an equal importance for each image.

The area under the receiver operator characteristic curve (AUROC) considers the sensitivity and specificity for a given positive class c over all thresholds of the model's predicted probabilities (Fig. 2.3c). The area under the resulting ROC curve is a commonly reported metric in skin lesion classification studies [41, 54, 70]. As the AUROC curve considers all decision thresholds, this metric is sensitive to the predicted probabilities.

One challenge that arises when comparing the performance of humans and machines is that humans, in general, report a single discrete *prediction*, while machines give a *probability* distribution. Specifically, the predicted label $\hat{y}^{(i)}$ of the i-th lesion is defined as the most probable label within the predicted probability distribution $p^{(i)}$,

$$\hat{y}^{(i)} = j^* = \underset{j \in \{1,\ldots,C\}}{\operatorname{argmax}} \, p_j^{(i)} \tag{2.9}$$

where C is the number of classes, and $p_j^{(i)}$ is the j-th class probability of the i-th lesion.

When computing the AUROC curve for a human, the sensitivity and specificity of the predictions are used and the ROC curve is assumed to be linear (see Fig. 2.3c). In contrast, the probabilistic outputs of machines often produce non-linear ROC curves (e.g., [54]). In addition, while the ROC curve demonstrates the *limits* of the model's ability to discriminate [179], this considers all possible probability thresholds, rather than the actual predictions. Thus a probabilistic model that makes incorrect predictions can still achieve a high AUROC score. Further, in a multi-class scenario where the non-positive classes are all considered negative, a ROC curve may be influenced by class imbalances [56].

While other metrics, such as average precision, are used [46, 66], they are reported less frequently in the literature. In order to compare human and machine predictions over multi-class datasets across a variety of works, this chapter focuses on reviewing experiments within studies where diagnostic accuracy can be inferred. Limitations when relying on diagnostic accuracy are discussed in Section 2.4.1.

2.2 Machine Skin Disease Classification

This chapter primarily focuses on skin disease classification and discuss other tasks (e.g., segmentation) in the context of supporting classification. A classification system is seen as a pipeline or model ϕ and parameters θ of ϕ, and generally requires a dataset of the observable input data x (e.g., images, patient history), and, for training or evaluation, the desired output data y (e.g., disease diagnosis). To design a classification system, a general optimization is done,

$$\phi^*, \theta^* = \underset{\phi,\theta}{\operatorname{argmin}} \, E(\phi(x;\theta), y) \tag{2.10}$$

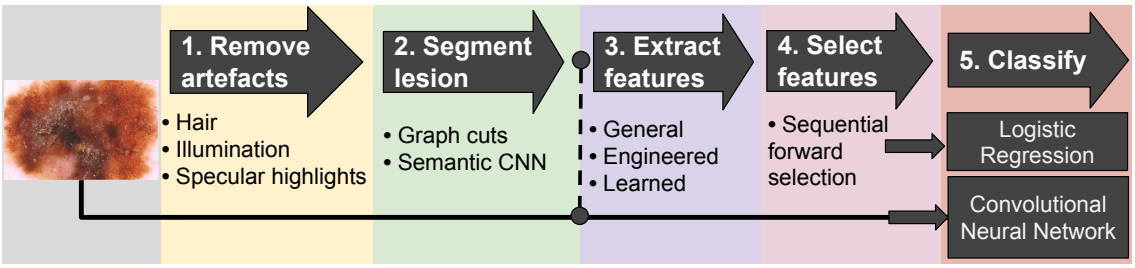


Figure 2.4: A common general pipeline to classify skin diseases. Image features can be extracted, then sent to a classifier (e.g., logistic regression). CNNs can extract features and classify diseases directly from images, but may also be used to extract features.

where $\phi(\cdot)$ is a model or pipeline that transforms the input data x into a predicted output $\hat{y} = \phi(\cdot)$, such that $\hat{y}$ matches the desired output y. θ are the parameters for the model/pipeline; $E(\cdot)$ measures the error between the predicted $\hat{y}$ and true output y, and can encode prior knowledge about the output or model parameters (e.g., regularization). ϕ^*, θ^* are the found model/pipeline and parameters, respectively, that minimize (globally or locally) $E(\cdot)$. There are many ways to optimize Eq. 2.10 as can be found in recent surveys [15, 27, 33, 127, 132, 134, 150]. A common approach is for a human to design a fixed model/pipeline ϕ, and to learn the parameters θ from the data using an explicit optimization (e.g., gradient descent). In the following sections, we discuss common pipelines/models, parameters, and optimization approaches, where each proposed component can be thought to be part of the general optimization of Eq. 2.10.

2.2.1 Sequential Pipeline Approach

Celebi et al. [35] proposed the following general pipeline ϕ to classify dermoscopy images as either benign or melanoma: 1) segment the skin lesion; 2) extract colour and shape-based features from the lesion border and regions within the lesion; 3) select a subset of discriminative features; and 4) use a machine learning classifier to distinguish among the classes. Related to Eq. 2.10, the pipeline and design choices ϕ are chosen by the authors (e.g., types of colour features to extract [115]), while the learned parameters θ are explicitly optimized by a support vector machine classifier (SVM).

This is referred to as a *sequential pipeline* approach since it follows a series of well-defined steps, where the fixed output from one step becomes the input to another (Fig. 2.4).

Many existing works propose a variation on this sequential pipeline, where a step may be improved or omitted. For example, Ballerini et al. [11, 12] used 960 clinical images from the Dermofit dataset (Sec. 2.1.4) to classify among five types of skin lesions with a 74.3% accuracy. They segmented lesions using a region-based active contour approach, extracted human engineered colour and texture features from the lesion and healthy skin separately, and selected features using sequential forward feature selection [80]. A hierarchical k nearest neighbour classifier clusters the images into two high-level classes (benign vs. pre-malignant

and cancer), followed by another classification to determine the sub-classes. Leo et al. [105] evaluated over 1,300 images of Dermofit composed of 10-classes, and followed a similar approach to achieve a classification accuracy of 67%. Shimizu et al. [145] segmented lesions and removed artefacts using a colour thresholding based approach. They extracted 828 colour and texture features based on the sub-regions of the segmented lesions, applied feature selection to select a subset of discriminative features, and used a two-stage hierarchical linear classifier to classify among four conditions.

2.2.2 Features Designed for Specific Dermoscopic Criteria

Instead of general colour and texture-based image features, some works specifically design features to capture known salient properties of a skin condition. This is common in melanoma classification, where the presence of specific dermoscopic criteria suggests melanoma (Section 2.1.3). For example, in order to detect blue-white veils in dermoscopy images, Madooei et al. [116] matched the lesion colours to a template of common blue-white veil colours. To detect and classify the types of streaks within dermoscopy images, Mirzaalian et al. [125] used a filter designed to capture the tubular properties of streaks. They segmented lesions using graph cuts [23], and used features derived from the detected streaks to train an SVM to distinguish the type of streaks. Fabbrocini et al. [55] designed separate pipelines and engineered features to classify seven dermoscopic criteria. For example, to classify irregular streaks, they segmented the lesion and compared the irregularity at the border to a reference threshold.

2.2.3 Learned Features of Dermatological Images

Rather than general engineered or features designed to target specific dermatological criteria, features can be learned from the data. In order to classify melanoma from non-melanoma in dermoscopy images, Codella et al. [40] applied an unsupervised sparse coding approach [117] to learn a sparse number of patterns that minimized an image reconstruction error. They also passed dermoscopy images into a CNN that was trained over the natural images (e.g., cats and dogs) of ImageNet [140], and extracted the CNN responses from select layers to use as feature vectors. They found that using these learned features to train an SVM gave a similar level of classification performance when compared to the previous state-of-the-art approach of using an ensemble of general engineered features. Over the clinical images of Dermofit (Sec. 2.1.4), Kawahara et al. [85] found that training a logistic regression classifier on features extracted from a pretrained CNN outperformed previously published approaches that relied on the classical pipeline and general engineered features.

Learning features directly from the images can also simplify the overall pipeline (Fig. 2.4) as this approach does not rely on engineered image features that require careful lesion segmentation (e.g., computing features at the border of the lesion). Avoiding lesion segmentation may be desirable as segmentation is challenging [36]. For example, 16% of the lesions

segmented by the top performing lesion segmentation method of the 2017 ISIC challenge had a Jaccard Index of less than 0.6, which is considered a failure [41]. These segmentation errors may propagate to errors in the features, which may decrease classification performance.

2.2.4 Joint Optimization to Learn Features and Classify

The approaches described so far extract fixed features from the images $f(x)$, and perform a separate optimization for classification,

$$\theta^* = \underset{\theta}{\operatorname{argmin}} E\left(\phi(f(x);\theta), y\right) \tag{2.11}$$

where $\phi(\cdot)$ is a user specified classifier (e.g., SVM, logistic regression classifier) parameterized by θ. The parameters learned when optimizing Eq. 2.11 are based on the fixed (possibly learned) set of image features $f(x)$, under the assumption that they will prove useful for classification. This section looks at works that combine feature learning and classification in a single optimization.

Deep learning [102] involves training a model composed of stacked layers of trainable parameters that learn non-linear feature representations of the data. Deep learning is widely used in skin lesion analysis, with the organizers of the 2017 ISIC skin challenge [41] (Section 2.1.4) noting that among the entries of this public challenge:

> "All top submissions implemented various ensembles of deep learning networks."

One type of deep learning model well suited for image classification is the CNN. The structure of the CNN considers the properties of images (locality of features, spatial invariance) and learns to transform the image pixels into discriminative feature representations. As all parameters within the CNN are learned, a CNN can be thought of as "synthesizing their own feature extractor" [103].

In this approach, a human designed CNN architecture $\phi(\cdot)$ is chosen, and an explicit optimization algorithm finds the CNN's parameters,

$$\theta^* = \underset{\theta}{\operatorname{argmin}} E(\phi(x;\theta), y) \ . \tag{2.12}$$

In contrast to Eq. 2.11, this equation does not have the human chosen representation of the features (i.e., $f(\cdot)$ in Eq. 2.11). Rather, the parameters to compute the features, and the parameters to classify are learned within a single optimization.

There are many possible error functions $E(\cdot)$, but a common choice for classification (others discussed in Sec. 2.2.4) is the categorical cross-entropy loss function,

$$E(p, y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_j^{(i)} \log(p_j^{(i)}) \tag{2.13}$$

where N is the number of images, C is the number of classes (e.g., types of skin diseases), $y_j^{(i)}$ is a one-hot encoded ground truth label, and $p_j^{(i)}$ is the predicted j-th class probability for the i-th image. Stochastic gradient descent can be used to learn the parameters θ that minimize Eq. 2.12, where the parameters θ of the model ϕ are iteratively updated [103].

CNNs for Classifying Skin Diseases

Many works (e.g., [90, 122, 113]) that use CNNs to classify skin diseases rely on CNN architectures (e.g., VGG16 [147]) that perform well when classifying natural images (e.g., ImageNet [140]). The parameters of the CNN learned over the natural images are stored (referred to as a *pretrained* CNN) and are used to initialize the weights of the CNN before training on a different target domain, such as skin images. The process of refining the learned parameters to a new target domain is referred to as *transfer-learning* or *fine-tuning* the CNN.

While several CNN approaches ignore lesion segmentations [90, 122, 113], which simplifies the overall pipeline (Sec. 2.2.3), explicitly localizing the skin lesion prior to training a CNN may reduce distracting background artefacts and improve overall performance. Yoshida et al. [175] trained a CNN to classify melanoma from nevi using dermoscopy images, where the major axis of each lesion was aligned in order to better capture the lesion asymmetry that is commonly associated with melanoma. A CNN trained using image augmentations that were constrained to maintain this alignment outperformed a CNN trained on non-aligned lesions when the same amount of image augmentation was performed. Yu et al. [176] used a two-step process where the lesion is first segmented using a fully-convolutional neural network trained to segment skin lesions, then the lesion is cropped based on this segmentation and passed to a CNN for classification. Using this approach, Yu et al. [176] ranked first place on the ISBI-ISIC 2016 skin lesion classification challenge (Sec. 2.1.4).

Although deep neural networks are often used to classify skin lesions [41], not all groups report better performance when compared to using a separate feature extraction and classification approach (Eq. 2.11). Sun et al. [151] collected 6,584 clinical and dermoscopy skin images, spanning 198 classes of common skin diseases from an online source. They trained an SVM on general engineered features and achieved a classification accuracy of 52.19% over the 198 classes, outperforming the 50.27% accuracy achieved using a CNN (VGG16 [147]) pretrained over ImageNet and fine-tuned to classify the skin conditions. A similar result was also found by Yang et al. [174].

Ge et al. [63] represented skin images as concatenated l2-normalized responses from ResNet-50 [74] and VGG16 [147] fine-tuned on skin images. They extracted features using compact bilinear pooling [62], and trained an SVM to classify among 15 types of skin diseases using skin lesions acquired as both a clinical and dermoscopy image, outperforming a single fine-tuned CNN. They summed the predicted probabilities from each imaging modality

together to form a final prediction accuracy of 71%. They used 24,182 training images and 8,012 testing images from an internal dataset known as "MoleMap".

Other Classification Loss Functions

Cross-entropy is a common loss function used to train a CNN (Eq. 2.13); however, other losses are also used. Ge et al. [64] incorporated clinical and dermoscopy images into a single CNN model trained to minimize the mean squared loss between the predicted and true vectors, and reported only minor differences in overall performance when compared to the cross-entropy loss. They used class activation maps [178] to find salient areas of the image, and extracted dense features from the diseased area using bilinear pooling [62]. They achieved a classification accuracy of 70% accuracy over 15-classes.

Demyanov et al. [46] trained a CNN using a tree-loss function that incorporated a human defined skin disease taxonomy. This taxonomy allows data to be labeled with different granularities. For example, a lesion could have the general label of "benign" (coarse granularity) and the more specific label of "blue nevus" (fine granularity). Using an internal dataset of 40,173 dermoscopy and clinical images, composed of 15 skin conditions, they trained ResNet-50 [74] using their tree-loss function and obtained 64.8% accuracy, demonstrating a small but consistent improvement to accuracy when compared to training without the tree-loss function.

2.2.5 Incorporating Non-Visual Information

While our focus in this survey is on visual classification, other non-visual information may provide important context when classifying skin diseases. Razeghi et al. [136] collected answers that humans gave to 37 simple questions about skin images (e.g., is the patient an infant, child, or adult?), as well as extracted general engineered features from the images. Using 2,309 clinical images from an online source composed of 44 disease types, they manually placed a bounding box around the lesion in the image, and trained a random forest to classify the skin diseases. Using only visual information, a trained random forest yielded 15.76% accuracy. Using only the human given answers to questions yielded 16.58% accuracy, and combining both yielded 25.12% accuracy. Kawahara et al. [91] incorporated clinical images, dermoscopy images, and patient meta-data (e.g., lesion location, sex) in a single CNN model designed to jointly classify multiple types of dermoscopic clinical criteria (e.g., type of streaks) and skin disease diagnoses. This approach reached an average classification accuracy of 73.7% when classifying skin diseases and dermoscopic criteria, which was an improvement over training on a single modality.

2.2.6 Image Retrieval

Adopting machine diagnoses into clinical practice may be hindered if the model does not offer an intuition into how the diagnoses are made. One approach towards more interpretable models is to retrieve images of known diseases that are visually similar to a user's lesion, allowing a user to visually inspect similar images of known diseases and infer a diagnosis.

Given a test query image q, the goal of image retrieval is to find the image $x^{(i)}$ within a dataset of known skin diseases that is most similar to the query image q. The corresponding known label $y^{(i)}$ is used as the prediction $\hat{y}^{(i)}$ for the unknown query image,

$$x^{(i)}, y^{(i)} = \underset{i \in \{1,\ldots,N\}}{\operatorname{argmin}} D(f(q), f(x^{(i)})) \tag{2.14}$$

where N is the number of samples in the labelled dataset, $f(x^{(i)})$ computes the features for the i-th image of the labelled dataset, and $D(a, b)$ measures the dissimilarity (e.g., cosine distance) between two feature vectors.

A variation on this approach is to find the $k > 1$ most similar images and a corresponding ranked list of diagnoses. Ballerini et al. [13] extracted colour and texture features from skin lesions, selected and combined features using a genetic algorithm, and retrieved up to $k = 10$ labelled images that had the lowest distance $D(\cdot)$ (e.g., euclidean distance) in feature space $f(\cdot)$ to a given query image. Bunte et al. [29] proposed an image retrieval system that retrieved dermoscopy images of similarly coloured lesions. They extracted colour-based features from manually selected patches within the lesion and healthy skin, learned features based on four classes of colours, and retrieved images using a k-nearest neighbourhood, where k ranged from 1 to 25.

Kawahara et al. [86] used a minimal path approach to find a progression of visually similar images between two query images. This may be useful in finding images related to disease progression (e.g., from benign to malignant). Skin lesions were represented as nodes in a graph, with edges representing the visual dissimilarity between lesions in a feature space based on the responses of a pretrained CNN.

Kawahara et al. [91] fine-tuned a pretrained CNN to classify both the disease and the 7-point checklist criteria [9]. They used the CNN layer responses $f(\cdot)$ to represent images and retrieved the class from the image with the lowest cosine distance $D(\cdot)$ to a query image. Over five classes of skin diseases, they reported an averaged retrieval accuracy of 71.1%. Tschandl et al. [159] used a similar approach for dermoscopy images and found that image retrieval had comparable accuracy with classification and allowed for better recognition of diseases that occur in datasets that the CNN was not trained on.

2.3 Dermatologist and Machine Performance

This section examines works that report the skin lesion classification performance of human dermatologists and machines from the same dataset. Dermatology studies that report lesion diagnosis via static images are also reported.

2.3.1 Ground Truth for Dermatologists' Classifications

Studies that measure human performance have dermatologists make diagnoses based on the provided static information (e.g., images, curated patient history). These diagnoses are compared to the "ground truth" class labels, which are determined by more rigorous diagnoses procedures. These procedures vary, but often consists of diagnoses by histopathology, a consensus of experts, interactive face-to-face sessions between dermatologists and patients, or a combination of approaches [162].

2.3.2 Dermatologists Compared with Machine

Ferris et al. [58] manually segmented skin lesions, extracted shape, colour, and texture based features, and trained a decision forest over 273 dermoscopy images, and tested the classification performance on 40 benign and 25 malignant dermoscopy images. Over the same test dataset of 65 lesions, 30 clinicians had an averaged melanoma sensitivity of 70.8% and specificity of 58.7%, whereas the automatic classifier had a melanoma sensitivity of 96% and specificity of 42.5%.

Codella et al. [42] and Marchetti et al. [118] compared an ensemble of top performing machine classification approaches, which included CNNs, to the average of eight dermatologists. Over 100 dermoscopy test images, the automated system achieved a higher accuracy (76%) than dermatologists (70.5%) when classifying 50 melanoma from 50 benign neoplasm images [42]. The eight dermatologists achieved an averaged sensitivity of 82% and specificity of 59%, while five top performing automated approaches on the ISIC-2016 challenges achieved a voting average sensitivity of 56% and specificity of 90%.

Esteva et al. [54] collected a dataset of 129,450 clinical images, which included 3,374 dermoscopy images, and spanned across 2,032 types of skin diseases. They grouped classes together based on their clinical similarity as per a human-defined taxonomy, which yielded 757 partitions (classes) for training. They used the Inception-V3 architecture [153], pre-trained over ImageNet [140], and fine-tuned the model on the partitioned classes. They reported results over different levels of the taxonomy, where the sum of the predicted probabilities in the descendant nodes determined the higher level classes predicted probabilities. Over a nine-class partition of dermatologist inspected images (i.e., may not be verified via histopathology), the CNN achieved an overall skin disease classification accuracy of 55.4%, which is comparable to the accuracy achieved by two dermatologists (53.3% and 55.0%).

To compare a CNN with humans in similar training conditions, Tschandl et al. [160] showed 298 dermoscopy images from six different lesion classes to 27 medical students (without prior dermoscopy experience), and provided only the corresponding diagnosis of each image without explaining the diagnostic features. The same images were used to fine-tune an Inception-V3 CNN architecture [153] pretrained over ImageNet, where the last layer was replaced to match the target classes. Both the students and CNN then diagnosed the diseases from 50 test images. The CNN achieved a diagnostic accuracy of 69% and was reported to demonstrate a similar diagnostic agreement as the average agreement among students. When diagnosing malignant lesions (basal cell carcinoma and melanoma) from benign, the CNN achieved a higher sensitivity (90% for CNN, 85% for students), but lower specificity (71% for CNN, 79% students) than the students' average scores.

Han et al. [71] formed a dataset of 49,567 hand and foot nail images by using manually labelled data, assisted by first training a hand and foot CNN classifier, followed by a region-CNN [137] trained to localize the nail plate, and an image quality CNN that eliminates poor quality nail images. They showed that a CNN could classify nail images that contain onychomycosis (a nail fungal infection) from other nail disorders with a higher Youden Index (sensitivity + specificity - 1) (67.62%) than then the average of 42 human dermatologists (48.39%) over 1133 images.

Han et al. [70] fine-tuned a pretrained CNN (ResNet-152 [74]) on 19,389 manually cropped clinical images taken from primarily an Asian population (Asan dataset). The training dataset was composed of 248 classes of skin diseases, while testing was done on an aggregated 12-class subset. The CNN tested over images from an Asian population achieved an accuracy of 57.3%, and 55.7% over the 12-classes of the Asan dataset, and the 10-classes of Dermofit (Sec. 2.1.4), respectively. Additional experiments comparing the diagnoses of 16 dermatologists over a subset of these images had, in general, a ROC curve inside the ROC curve produced by the CNN.

Yang et al. [174] had general doctors, junior dermatologists, and expert dermatologists classify skin images from 198 classes of skin diseases. Two doctors from each category were invited to independently classify images and discuss the diagnosis when they differed. The accuracy was 49.00% for general doctors, 52.00% for junior dermatologists, and 83.29% for expert dermatologists. The accuracy of the top performing CNN was 53.35%, which was lower than the expert dermatologists, but comparable with general doctors and junior dermatologists.

Haenssle et al. [67] trained a CNN to classify dermoscopy images as either a benign nevi or melanoma using training images from a variety of sources. Using 100 test dermoscopy images, they compared the classification results of the CNN with 58 dermatologists. On average, dermatologists had a sensitivity of 86.6% and specificity of 71.3%, while a CNN tested over the same images achieved a sensitivity of 95% and specificity of 63.8%.

Fujisawa et al. [61] fine-tuned a pretrained CNN on clinical images to classify among 21 disease classes and aggregated the predicted classes within a skin tree hierarchy. Using the diagnoses aggregated at the third level of the tree with 14 classes, the CNN achieved an accuracy of 76.5%, outperforming the averaged diagnostic accuracy of 13-board certified dermatologists (59.7%) and nine dermatology trainees (41.7%).

Tschandl et al. [161] combined the predictions from a CNN trained on dermoscopy and a CNN trained on clinical close-up images to form a final diagnosis. When compared with 95 human examiners with varying levels of expertise [148], the CNN had a higher number of correct specific diagnosis (37.6%) than the human examiners (33.5%), but lower than human expert dermatologists (40.0%).

2.3.3 Comparing Dermatologists on Static Images

To better estimate human performance, this section primarily examines *store-and-forward* teledermatology studies, where the patient data (e.g., lesion image, patient history) is sent to a dermatologist for review [158],

To compare how different types of static images influences human performance, Sinz et al. [148] had 95 human examiners (including 62 dermatologists) classify 50 images randomly sampled from 2,072 cases into one of 51 possible diagnoses. Using clinical images, the averaged accuracy was 26.4%. Using dermoscopy images, the averaged accuracy improved to 33.1%, indicating that performance depends on the imaging modality.

To compare *in vivo* diagnosis and diagnosis via static images, Carli et al. [30] collected 256 lesions composed of seven classes of biopsy verified diseases. Using the consensus of two dermatologists (in disagreement, a third dermatologist was consulted), they reported a diagnosis accuracy of 40.1% during clinical examinations without dermoscopy. When *in vivo* dermoscopy was incorporated with the clinical examination, the accuracy improved to 72.3%. The accuracy dropped to 54.7% when the dermatologists had access to only the dermoscopy photographs and patient history, but not clinical information.

Weingast et al. [169] had 263 patients photograph their own lesions, when possible, using a mobile camera, and provide additional questionnaire information. They collected a wide variety of skin conditions, which were typical of the authors outpatient unit. Multiple teledermatologists reviewed each case, and overall, 49% of the gathered cases could be correctly diagnosed via teledermatology when compared to a face-to-face consultation (a differential diagnosis was allowed in some cases i.e., top-2 accuracy Eq. 2.1). The teledermatologists reported only 61% of the cases contained sufficient information to make a diagnosis.

In a prospective study with 63 dermoscopy images, Walker et al. [167] used a CNN to extract visual feature representations that were converted into sound and visually or audibly analyzed by humans to detect cancerous skin lesions, achieving a sensitivity of 86% and specificity of 91%.

Brinker et al. [26] had 157 dermatologist assess 100 dermoscopy images and 145 dermatologists assess 100 clinical images composed of nevi and melanoma skin lesions. Dermatologists provided a management decision (biopsy vs. reassure patient), achieving an average of 74.1% sensitivity and 60.0% specificity for dermoscopy images; and, 89.4% sensitivity and 64.4% specificity for clinical images.

2.4 Discussions

This section lists the challenges of comparing across studies, summarizes the reported performance of selected human and machine skin disease classification works, and discusses potential limitations and sources of error within image-based diagnoses of skin conditions.

2.4.1 Challenges of Metrics and Comparing Skin Studies

This chapter primarily focused on the metric of diagnostic accuracy as it is commonly reported (or can be inferred) in both clinical and computer vision studies, and it gives us a single intuitive metric for multi-class problems. However, relying on diagnostic accuracy assumes that all errors are equal, which may hide a poor performance on infrequently occurring diseases. Other metrics, such as averaged diagnostic sensitivity or precision, address the class imbalance problem by giving an equal weighting to each class, resulting in a higher weighting of infrequent conditions. All these metrics are limited since clinically, some conditions are more important to correctly diagnose than others (e.g., a false-negative melanoma diagnosis can be fatal). One potential solution is to weight each misdiagnosis to account for the severity of a misdiagnosis. However, establishing such a clinical weighting is non-trivial for multi-class problems, and would require significant expert knowledge.

Another approach is to ignore diagnostic performance and instead focus on predicting appropriate treatments (e.g., [148]). While this considers the clinical implications of a disease, it requires a consensus on appropriate treatments, which may change as new treatments become available. Another complication is illustrated in the case of melanoma, where images that are biopsy verified are, by definition, ones that a dermatologist recommended for biopsy. Thus, biopsy images labeled as benign are clinically suspicious enough that an expert flagged them for biopsy. One may question if the goal of machine classification should be to replicate the dermatologist's decision or to classify the underlying disease.

A limitation in comparing across studies is that the difficulty of diagnosing diseases depends on the dataset (e.g., some diseases display more consistent morphology), making it unclear if one particular methodology performs better of if the differences are due to the datasets. Studies that compare the performance of both humans and machine (Sec. 2.3.2) often compare over the same dataset, allowing for a fairer comparison.

Nevertheless, with these limitations stated, diagnostic accuracy is used as our primary metric, largely due to insufficient information provided in many studies to infer other metrics

and the challenges associated with choosing a single more descriptive metric. Results are aggregated across different studies, composed of a variety of datasets, in order to compare the performance of humans and machines.

2.4.2 Comparing Human and Machine Performance

Table 2.5 shows 44 skin condition classification experiments selected from 21 studies, spanning both clinical and computing research. The works in this table were selected based on the following criteria: 1) they compared humans and machines over the same dataset, or 2) they reported human and machine performance separately on a multi-class (greater than two) dataset. Experiments where the *predictions* (Eq. 2.9) of a model could not be inferred were omitted. This primarily occurred when only the AUROC scores were reported.

As not all studies report accuracy, accuracy was inferred given the other reported metrics. Occasionally, the exact sensitivity and specificity were not given, and these were estimated from the reported graphs. For studies that had predictions made by multiple humans, the accuracy was computed using the average human performance.

Table 2.5 reports the dataset and the number of images used to evaluate, the input modality, whether human or machines did the diagnosis, the number of classes, and the accuracy over the entire test set. Fig. 2.5 plots the number of classes versus the reported accuracy, separated based on machine and human skin disease classification performance. A general trend is observed, where as the number of classes increases, the accuracy decreases.

Similar accuracy is found when averaged across studies for both humans and machines (Fig. 2.6). As well, Fig. 2.7 highlights that the inclusion of patient history (e.g., questionnaire, age, sex) yields small changes to accuracy, with the exception of one non-deep learning study that included 37 user supplied answers [136].

2.4.3 Limitations and Sources of Errors in Image-Based Diagnosis

When developing an image-based classification system, there are several limitations and sources of potential errors. The "ground truth" disease labels may have errors, even when confirmed via histopathology. Monheit et al. [128] found that due to conflicts in the expert histopathology diagnoses, 8.8% of lesions required more than two histopathological evaluations before reaching a final diagnosis. Elmore et al. [53] collected 240 biopsy cases and used the consensus of three human experts to label each case into one of five categories that indicated a progressively increasing melanoma risk. These consensus labels were compared to diagnoses given by 187 pathologists, and the authors found that the three diagnoses categories spanning "moderately dysplastic nevi to early stage invasive melanoma were neither reproducible nor accurate" [53].

Patients may be limited by their ability to capture high quality images. Weingast et al. [169] had patients attempt to acquire an image of their own lesion using a mobile phone

Table 2.5: Selected skin disease classification approaches and diagnostic performance. *N.Images* indicates the number of images in the dataset. *N.Test* indicates the number of images used to test (includes cross-validation). *Derm.*, *Clinic.*, and *Meta.* indicate dermoscopy images, clinical images, and non-image patient history (meta-data), respectively, where a star (*) indicates *in vivo* data. *H.vs.M* indicates if the diagnosis was made by a human or machine. *Acc.* indicates diagnostic accuracy.

	Year	Dataset	N.Images	N.Test	Derm.	Clinic.	Meta	H.vs.M	Classes	Acc.
[58]	2015	Internal	-	65	✓			human	2	63.35
[58]	2015	Internal	273	65	✓			machine	2	63.08
[42]	2017	ISIC-100	-	100	✓			human	2	70.50
[42]	2017	ISIC-100	1000	100	✓			machine	2	76.00
[67]	2018	Internal	-	100	✓			human	2	74.40
[67]	2018	Internal	-	100	✓		✓	human	2	78.30
[67]	2018	Internal		100	✓			machine	2	81.60
[71]	2018	Asan	-	1133		✓		human	2	75.80
[71]	2018	Asan	49,567	1133		✓		machine	2	80.00
[26]	2019	ISIC-100	-	100	✓			human	2	62.82
[26]	2019	MED-NODE	-	100		✓		human	2	69.40
[54]	2017	Stanford	-	180	✓	✓		human	3	65.78
[54]	2017	Stanford	127,463	127,463	✓	✓		machine	3	72.10
[12]	2013	Dermofit	960	960		✓		machine	5	74.30
[91]	2018	Atlas	2018	395	✓	✓		machine	5	71.10
[91]	2018	Atlas	2018	395	✓	✓	✓	machine	5	73.70
[160]	2017	Internal	348	50	✓			human	6	74.00
[160]	2017	Internal	348	50	✓			machine	6	69.00
[30]	2002	Internal	-	256		✓*	✓*	human	7	40.62
[30]	2002	Internal	-	256	✓		✓	human	7	54.69
[30]	2002	Internal	-	256	✓*	✓*	✓*	human	7	72.27
[54]	2017	Stanford	-	180	✓	✓		human	9	54.15
[54]	2017	Stanford	127,463	127,463	✓	✓		machine	9	55.40
[105]	2015	Dermofit	1300	1300		✓		machine	10	67.00
[85]	2016	Dermofit	1300	1300		✓		machine	10	81.80
[70]	2018	Dermofit	20,689	1300		✓		machine	10	55.70
[70]	2018	Asan	19,389	1,276		✓		machine	12	57.30
[61]	2019	Internal	-	1260		✓		human	14	41.70
[61]	2019	Internal	-	1820		✓		human	14	59.70
[61]	2019	Internal	6009	1142		✓		machine	14	76.50
[46]	2017	MoleMap	40,173	1776	✓	✓		machine	15	69.10
[64]	2017	MoleMap	26584	7975		✓		machine	15	61.20
[64]	2017	MoleMap	26584	7975	✓			machine	15	61.30
[64]	2017	MoleMap	26584	7975	✓	✓		machine	15	70.00
[63]	2017	MoleMap	32,194	8,012	✓	✓		machine	15	71.00
[136]	2014	dermis	2309	1429		✓		machine	44	15.76
[136]	2014	dermis	2309	1429		✓	✓	machine	44	25.12
[148]	2017	Internal	-	2072		✓		human	51	26.40
[148]	2017	Internal	-	2072	✓			human	51	33.10
[151]	2016	SD-198	6,584	3292		✓		machine	198	52.19
[174]	2018	SD-198	-			✓		human	198	49.00
[174]	2018	SD-198	-			✓		human	198	52.00
[174]	2018	SD-198	-			✓		human	198	83.29
[174]	2018	SD-198	6584	3292		✓		machine	198	56.47

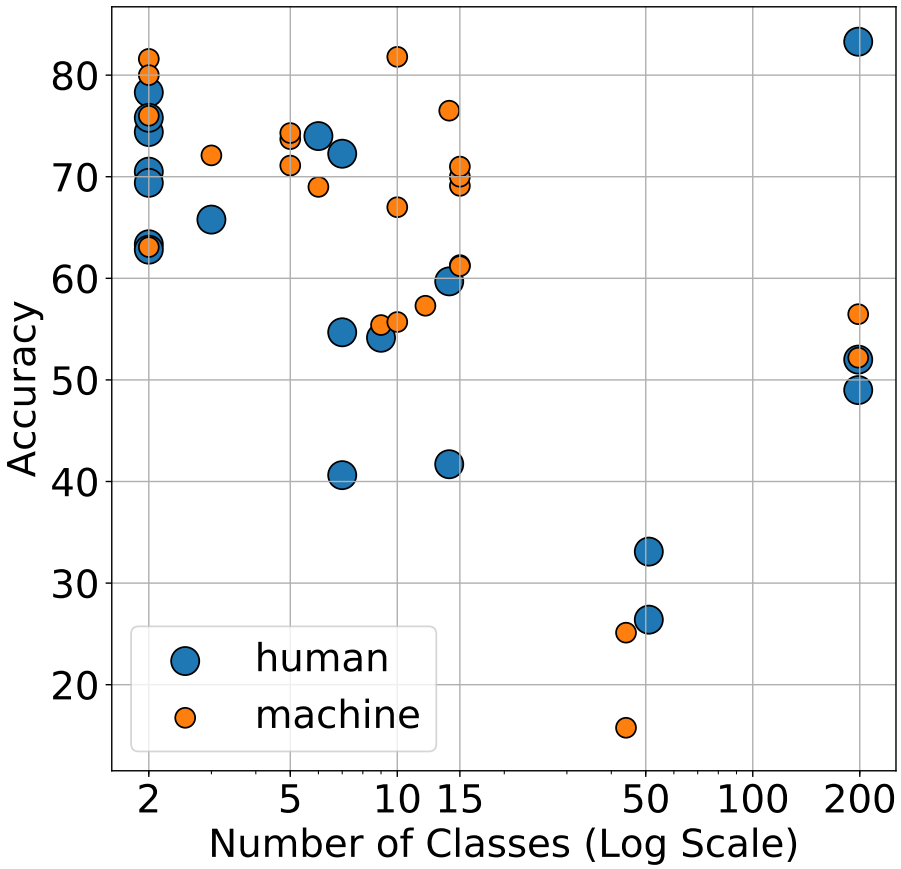


Figure 2.5: Skin disease classes versus reported model accuracy. Each coloured dot represents a experiment from Table 2.5, where the diagnosis was made by either a human or machine.

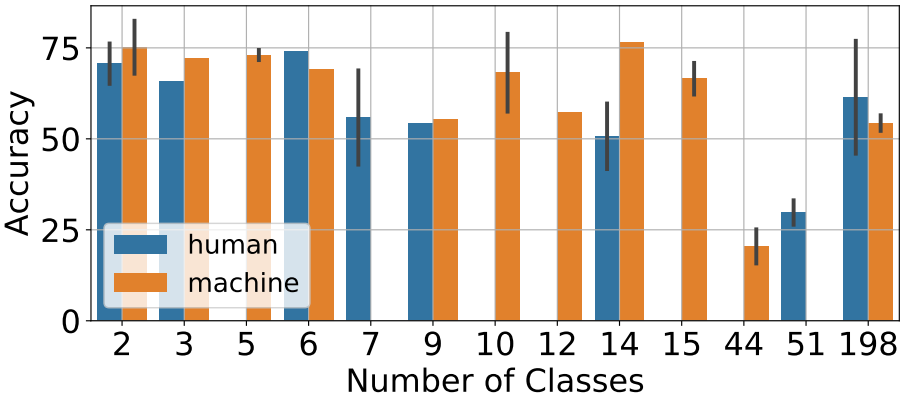


Figure 2.6: Averaged accuracy of the experiments in Table 2.5, grouped by the number of classes. On average, similar performance of both human and machines is reported.

camera. However, 81% of patients required assistance in acquiring images, partly due hard-to-reach lesions, and challenges in focusing and choosing an appropriate field-of-view. Even with assistance, 39% of the cases were reported to have insufficient information to make a

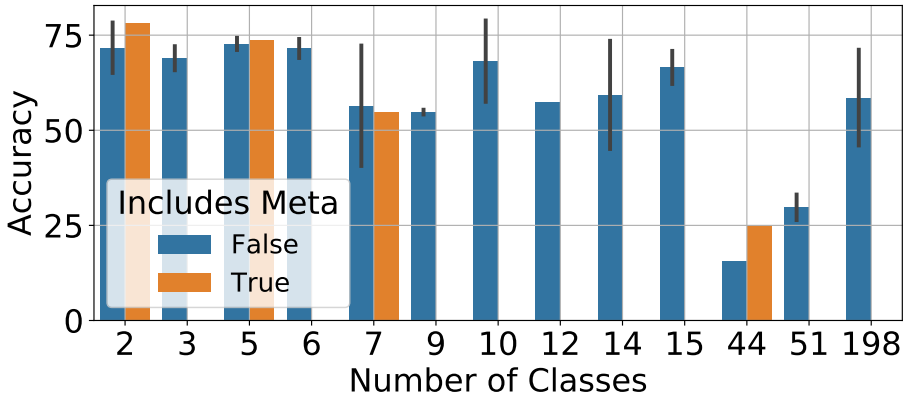


Figure 2.7: Averaged accuracy of the experiments in Table 2.5, grouped by the number of classes. Similar accuracy is reported when additional meta data (i.e., patient history) is included in the diagnosis.

diagnosis via teledermatology (but could be diagnosed face-to-face), indicating significant challenges in acquiring quality images.

Images may contain insufficient or misleading information easily resolved during a face-to-face examination. In Sec. 2.3.3, the reported differences in human diagnosis performance during teledermatology suggests that diagnosing via static images may be significantly more challenging than diagnosis during face-to-face consultations. A further example is given by, Hogan et al. [76] who documented a patient supplied image that appeared to contain serious complications, but on a face-to-face inspection revealed a crust covering a well-healing wound. Thus, claims that machines have reached human-level diagnostic ability should be considered in the context of static images.

The role non-visual information (e.g., patient history, questionnaire data) takes in the diagnostic procedure and what information should be gathered is not clear. Machine diagnoses systems that do utilize non-visual patient data report a mixed impact to performance, ranging from minimal [91] to substantial improvements [136]. Experienced dermatologists exhibit minimal improvements to diagnosis when given patient history (age, sex, body location site) in addition to an image, but those with less experience show a greater improvement with access to patient history [67]. Acquiring this data outside of face-to-face consultations may also be challenging. Weingast et al. [169] reported that most patients over 60 years needed assistance completing a computer questionnaire.

Another consideration is how transferable across datasets and populations the models are. Han et al. [70] report an accuracy of 55.7% over the 10-classes of Dermofit. This is significantly lower than other reported works that train and test only over Dermofit (e.g., 81% [85]). As Han et al. [70]'s model was trained on an Asian population and tested on a European population, this drop in accuracy may be due to the differences in how skin

diseases manifest across populations, or signify a lack of transferability in learned features across datasets due to image acquisition protocols, or both.

These sources of errors and limitations are potentially compounding, where ground truth errors in training, may compound with low quality acquired images, and a lack of model transferability across populations.

2.5 Conclusions

While there are still significant challenges in skin disease diagnosis, in 2017 dermatologists from a variety of institutions wrote the following statement [108]:

> "With the physician workforce projected to remain relatively flat, the specific ratio of dermatologists to population will decrease over time, especially in rural areas. These projections indicate a current and future challenge to ensure patient access to appropriate dermatologic care."

Automated analysis of skin conditions has the potential to alleviate the diagnostic requirements of dermatologists, making this a field worthy of investigation.

Given that recent studies report comparable accuracy performance when comparing dermatologists and machines, and considering the reported performance across independent teledermatological and machine studies, it is reasonable to conclude that machine accuracy is nearing the performance of human dermatologists in a teledermatological scenario. However, given the differences in performance when dermatologists diagnose via teledermatology [169], machine classification accuracy may be significantly lower than a face-to-face consultation with a dermatologist. As the diagnoses performance of general practitioners is reported to be twice as low as dermatologists [143], machine classification of skin diseases may have increased utility among general practitioners, who are often the first clinicians to examine dermatological disorders. We highlight that when humans classify among the 1000-classes of the natural images within ImageNet, the reported top-5 error (considers a match in any of the top-5 predictions to be correct) is 5.1%-12% [140]. While not directly comparable due to the different number of possible classes considered, the relatively low accuracy for humans classifying skin diseases commonly reported in Table 2.5 indicates the challenges of classifying skin diseases from images. Finally, we note that diagnostic accuracy, which is focused on in this chapter, gives us a limited understanding of performance, and does not consider the severity of misdiagnosing certain conditions.

Chapter 3

Deep Features to Classify Skin Lesions

3.1 Introduction

Skin cancers are commonly grouped into either melanoma or non-melanoma skin cancers. Melanoma skin cancers have a higher mortality rate, while non-melanoma skin cancers have a higher incidence rate. Early detection is important for treatment, which can differ based on the cancer type [3]. This makes systems to automatically classify types of skin lesions a potentially useful screening tool for initial referrals or as an additional supporting/safety-net expert system. As melanoma has a higher mortality rate than non-melanoma skin cancer, distinguishing between cancer and noncancerous melanoma skin images has attracted considerable research [99]. However, non-melanoma skin cancer is the most common cancer in light skin populations and, while it has a lower mortality rate than melanoma skin cancer, it places a large burden on quality of life and health care services [111]. Thus distinguishing among melanoma, non-melanoma, and other types of benign skin lesions are an important component of a practical skin diagnosis tool and is a focus of this work.

3.1.1 Non-melanoma Skin Cancers

Focusing on non-melanoma skin cancers, Ballerini et al. [12] used a hierarchical K-nearest neighbors based approach to classify among 5-classes of skin lesions using images captured with a colour camera. Leo et al. [105] extended this approach to classify 10-classes of skin lesions that contained both melanoma and non-melanoma as well as benign skin lesions. This dataset [12, 105] (Fig. 3.1) of non-dermoscopy images is publicly available and allows us to compare methods. Shimizu et al. [145] used a similar strategy to classify among 4-classes of dermoscopy skin lesions that included both melanoma and non-melanoma lesions. These approaches [12, 105, 145] followed the similar pipeline of: preprocess the image; segment the lesion; extract a candidate set of conventional image and shape features; select a class discriminative subset of these features; and, train a classifier.

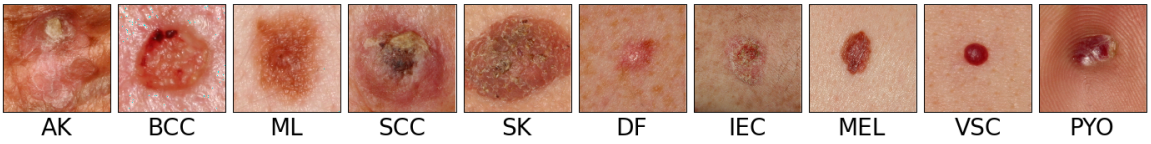


Figure 3.1: Dermofit images from each of the 10-classes randomly sampled. The first 5-classes make up the 5-class experiments.

3.1.2 Avoiding Skin Lesion Segmentation

In this work, we prefer to avoid lesion segmentations and complex preprocessing as these are non-trivial steps where errors can propagate (e.g., poor segmentations gives poor features) and may require subjective human intervention [36]. For example, the skin lesion segmentation approach of Li et al. [106] required manual initialization, post-processing, and utilized depth information. Further, they report large variations in the manual lesion segmentations done by dermatologists, which may indicate that segmentations are subjective. Thus, we focus on a state-of-the-art image feature extractor, that does not require lesions' segmentations nor complex preprocessing, in the form of a pretrained fully-convolutional neural network.

3.1.3 Pretrained Convolutional Neural Networks to Extract Features

CNNs have emerged as a powerful classification tool and are consistently used in competitions such as the ImageNet challenge, wherein researchers compete to classify hundreds of different natural objects [140]. CNNs not only give state-of-the-art results when trained for a specific task, but experiments have shown that the filters learned over the ImageNet dataset are generic and useful for other image tasks that the CNN was not originally trained for [40, 50, 147]. For example, Donahue et al. [50] used the AlexNet [100] architecture trained on ImageNet and found the responses from the first fully connected layer, FC6 (the sixth network layer), used to train a linear classifier outperformed conventional engineered features across a variety of benchmarked datasets of natural objects. Codella et al. [40] used pretrained CNNs to extract deep features from dermoscopy images to perform 2-class classification of two tasks (melanoma vs. non-melanoma and melanoma vs. atypical lesions) and found comparable performance with conventional low-level visual features.

3.1.4 Contributions

We convert a CNN into a fully-convolutional neural network (full-CNN) by converting the fully connected layers to convolutional layers [144]. This is an efficient approach to computing features over different spatial locations as it reuses the common convolutions done early in the network and can be used to compute features over multiple scales [144]. Aggregating features over the spatial dimensions has shown to improve predictive performance and the resulting feature vectors generalize well to other natural object image tasks [147].

Using a pretrained CNN as a feature extractor rather than training a CNN from scratch is attractive as it transfers learning (i.e., filters) from other domains where more training data is available, and shortens the time consuming training process. However, it is not obvious if the filters learned in a CNN trained on natural images will generalize well to those found on closeup skin images. Aside from recent dermoscopy work [40], most CNN-based works [50, 147] have focused on benchmarking similar natural objects. Thus we investigate whether CNN filters trained on natural objects generalize to multi-class (greater than two) classification of *non-dermoscopy* (i.e., without requiring a dermoscope) skin lesion images. We find that these features do generalize well and outperform previously published results over the same dataset, *without the aid of the corresponding lesion segmentations* used in previous approaches. We improve on the standard CNN as a feature extractor approach by using per-image normalization, a pretrained full-CNN to extract features from multiple scales, and by pooling across an augmented feature space, all of which yield classification improvements.

3.2 Methods

Given a skin image x with a corresponding class label y representing the skin lesion class, we want to extract image features $f = \phi(x)$ that differentiate well among the different class labels. To extract image features, we use the architecture of AlexNet [100] pretrained on the natural images found in ImageNet [140]. To extract features at multiple scales, we follow a similar approach to Sermanet et al. [144] to convert the CNN to a full-CNN. We convert the fully connected layers of AlexNet to convolutional layers, where these pretrained weights from the fully connected layers now act as convolutional filters. These filters can now be convolved with larger inputs (i.e., larger images) to efficiently extract responses at different scales. A skin image is passed through the network, and we extract the features from FC6 (now a convolutional layer) as FC6 has been shown to generally yield generic feature vectors [40, 50]. The responses from FC6 (i.e., the deep skin features) are used to train a logistic regression classifier to classify the skin lesions. We compute features at different scales by changing the size of the image. Thus, when the image is larger than the CNN's original receptive field, we get a feature vector with a height and width dimension, which corresponds to spatial locations in the larger input image. In order to reduce dimensionality and to achieve invariance to the spatial locations of the responses, we max-pool across the *spatial dimensions* (see Eq. 3.2) to get a single feature vector for the entire image.

3.2.1 Image Normalization and Preprocessing

Typically, images are normalized by subtracting the averaged activity over the training set to center the RGB values around zero [100]. As the CNN was trained over ImageNet images, we subtract from our skin images the averaged pixel activity of the ImageNet training images.

We also explore other normalization approaches. To provide some invariance to differences in lighting and skin tone, we hypothesize that subtracting the mean RGB pixel values computed over each *individual* image (per-image-mean) will improve the discriminative values in the resulting feature vector. We report results over different image normalization options in Table 3.1. Aside from resizing images, this is the only preprocessing we perform. We contrast this simple preprocessing to other competing approaches that require more complex preprocessing such as lesion segmentation, and specular highlight removal [12, 105, 145].

3.2.2 Pooled Deep Features for Augmented Images

A common approach to improve a CNN's classification accuracy is to augment the images [100, 135]. As skin lesions can potentially be imaged from a variety of camera rotations, we augment the images using a rotation by 0, 90, and 270 degrees as well as a left-right flip. Given the i-th image $x^{(i)}$, we augment and resize it to produce a j-th augmentation of the i-th image $\tilde{x}_j^{(i)}$. We normalize the augmented image and compute a feature vector by extracting the responses at FC6. For example, normalizing the image using the per-image-mean subtraction, we compute an augmented feature vector as,

$$\tilde{\boldsymbol{f}}_j^{(i)} = \phi(\tilde{x}_j^{(i)} - \mu(\tilde{x}_j^{(i)})) \tag{3.1}$$

where $\mu(x)$ returns the mean value for each colour channel in x, and $\phi(x)$ extracts the FC6 responses.

These augmented feature vectors (Eq. 3.1) could be used as additional samples to train a classifier and as additional image views during testing. However, there are additional time and memory costs associated with training and testing a classifier on more samples. Pooling across feature space creates a single representative feature vector for all the augmentations that allows us to keep the same time and memory benefits of having a single feature vector per image. We use a similar approach as [147], where instead of averaging across only left-right flips, we pool across N_{aug} augmentations. Combining the max-pooling over the full-CNN's spatial (height h and width w) dimensions with the mean-pooling in augmented feature space, we compute our augmented feature vector as,

$$\hat{f}_k^{(i)} = \frac{1}{N_{\text{aug}}} \sum_{j=1}^{N_{\text{aug}}} \max_{h,w} \left(\tilde{f}_{h,w,k,j}^{(i)} \right) \tag{3.2}$$

where $\max_{h,w} \left(\tilde{f}_{h,w,k,j}^{(i)} \right)$ computes the maximum spatial response of the k-th feature for the j-th augmentation of image i. These pooled augmented feature vectors summarize the augmentations of each image, while keeping the time and memory benefits of using fewer training/testing samples.

3.3 Results

We validate our approach on the Dermofit Image Library[1]. This dataset is composed of 1,300 skin images with corresponding class labels and lesion segmentations. There are 10 lesion categories (Fig. 3.1) in this dataset: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Melanocytic Nevus/Mole (ML), Squamous Cell Carcinoma (SCC), Seborrhoeic Keratosis (SK), Intraepithelial Carcinoma (IEC), Pyogenic Granuloma (PYO), Haemangioma (VSC), Dermatofibroma (DF), and Malignant Melanoma (MEL). As Ballerini et al. [12] report detailed results and experiments over 5 of these classes, we focus our comparison on these 5-classes (AK, BCC, ML, SCC, SK), but also benchmark over the entire 10-class dataset to compare with Leo et al. [105].

To divide our data, we follow the approach of Ballerini et al. [12], where we randomly split the dataset into three approximately equal sets with approximately the same distribution of class labels. We show cross validated results where two sets are used to train and one is held out to test. For ease of comparison, we report a single accuracy over all N (N=960 for 5-class and N=1,300 for 10-class) tested images, $\frac{1}{N}\sum_i^N \delta(\hat{y}^{(i)}, y^{(i)})$ where $\delta(\hat{y}^{(i)}, y^{(i)})$ returns 1 if both the predicted $\hat{y}^{(i)}$ and the true $y^{(i)}$ labels are equal to each other, else 0. For a fair comparison, we compute results for [12] with this measure of accuracy using their confusion matrix (Table 3.1 *row a*). To better indicate the performance per class, we report the confusion matrix across all classes. Following [12], we also report the results of grouping our 5-class predictions into a 2-class problem. Specifically, we group our 5-class predictions for BCC, SCC and AK together to form a *cancer and potential risk lesion* class and group our 5-class predictions for ML and SK together to form a *benign lesion* class.

In all experiments, we train a logistic regression classifier (using the default parameters) to classify the skin lesions using the deep features to represent each image. We use the Caffe [82] implementation of AlexNet [100] to extract the CNN pretrained feature vectors at FC6 (i.e., the sixth layer in the CNN).

3.3.1 Image Normalization

We start by examining the effect of normalizing the images prior to extracting features. We observe that on images resized to 227×227, subtracting the dermofit-mean-pixel (subtract the mean response over our Dermofit training images) and the per-image-mean (Table 3.1 *row c,d*) yield more accurate results than subtracting the ImageNet-pixel (Table 3.1 *row b*). Thus, we leave out the ImageNet-pixel subtraction approach from further experiments. We also experimented with subtracting the mean ImageNet and mean Dermofit images, but neither improved results. While we found improvements to the 5-class problem when compared with [12], for this experiment our 2-class accuracy is slightly lower.

[1] `https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html`

Table 3.1: Accuracy over all predictions. Rows with 1×1 indicate images of size 227×227 are convolved with the full-CNN and rows with 5×5 indicate images of size 339×339 were used. The plus sign (+) indicates concatenation of two feature vectors. The *aug* column indicates if image augmentation was used. The *norm* column indicates how images were normalized (e.g., subtracting the per-image-mean).

	method	aug.	norm.	5-class	2-class
(a)	[12]	✗	lesion seg.	75.1%	92.7%
(b)	1×1	✗	ImageNet-pixel	77.7%	90.6%
(c)	1×1	✗	dermofit-pixel	81.3%	92.1%
(d)	1×1	✗	per-image-mean	81.6%	92.2%
(e)	5×5	✗	dermofit-pixel	81.3%	93.1%
(f)	5×5	✗	per-image-mean	82.3%	91.9%
(g)	1×1	✓	per-image-mean	82.9%	93.0%
(h)	5×5	✓	per-image-mean	83.8%	94.7%
(i)	1×1 + 5×5	✗	per-image-mean	84.3%	93.0%
(j)	1×1 + 5×5	✓	per-image-mean	**85.8**%	**94.8**%
(k)	[105]	✗	lesion seg.	10-class = 67%	
(l)	1×1 + 5×5	✓	per-image-mean	10-class = **81.8**%	

The next set of experiments (Table 3.1 *row e,f*) use features computed at a higher resolution. Images are resized to 339×339 and 5×5×4096 dimensional feature vectors are extracted from the full-CNN. We max-pool across the spatial domain to get a single 4096 feature vector to train our classifier. We find that the per-image-mean subtraction works slightly better than the others and thus use it for the rest of the experiments.

3.3.2 Augmented Multi-scale Deep Features

We examine the effect of augmentation (Table 3.1 *row g,h*) by mean-pooling the augmented feature vectors across feature space (Eq. 3.2) and find this yields consistent improvements across both scales. We then examine the performance of feature vectors computed at two scales by concatenating the feature vectors (Table 3.1 *row i*), yielding further improvements to accuracy over a single scale. We concatenated the multi-scale feature vectors (instead of pooling) in order to capture differences in lesion scales as all images are taken at roughly 50 cm [105] from the skin. Concatenating the pooled-augmented feature vectors (Table 3.1 *row j*) yields the highest results in both the 2- and 5-class accuracy. We highlight that we improve accuracy to 85.8% over the previous results of 75.1% without using segmentations. We run our proposed approach over the full 10-class dataset without any further tuning (Table 3.1 *row l*) and find this generalizes well with an accuracy of 81.8%, outperforming the 67% accuracy reported by Leo et al. [105] (Table 3.1 *row k*).

These aforementioned experiments indicate that deep features do generalize well to these skin images and outperform competing approaches [12, 105], *despite our approach not using*

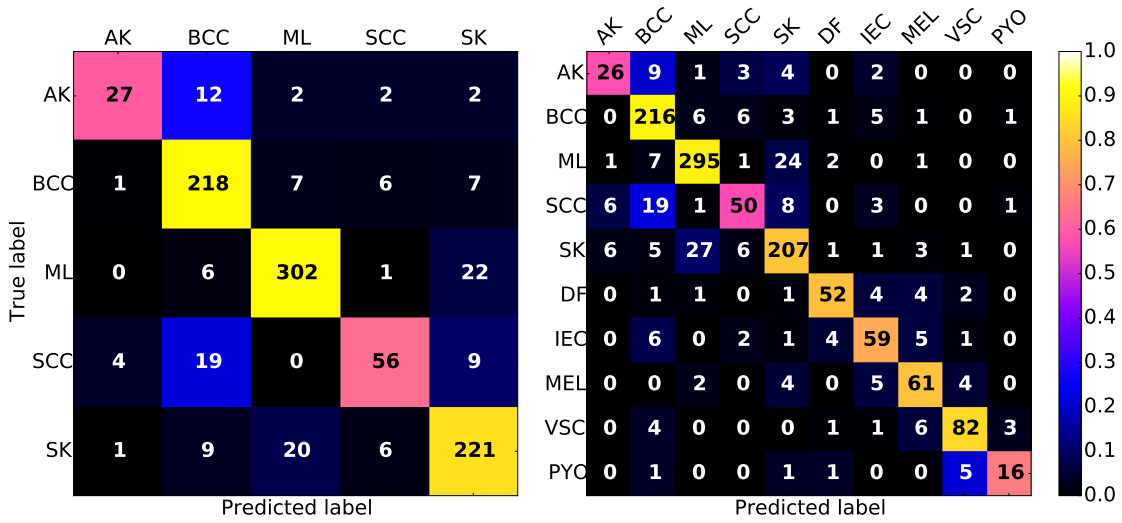


Figure 3.2: Confusion matrices for our proposed approach when trained for the 5-class (*left*) and 10-class (*right*) problem. Each cell shows the number of samples predicted to belong to each class. Colours show per-class accuracy values normalized across rows by the number of images in each class.

(nor requiring) any lesion segmentations. We found this result surprisingly remarkable as the pretrained CNN was optimized for natural images with considerably different appearance than closeup skin lesion images. A similar result was also found in the recent work of Codella et al. [40] (who reported *2-class* results over a *dermoscopy* dataset, in contrast to our *10-class* results over a *non-dermoscopy* dataset), and our findings further confirm the generalizability of pretrained CNNs to the skin domain (as opposed to work showing generalizability on more natural images [50, 147]).

We note that reporting accuracy over all images hides some large improvements to those classes with a small number of images. In particular, previous work reported 15.6% accuracy for the AK class where here we improve it to 60%. The confusion matrix for our full approach is shown in Fig. 3.2 showing a breakdown of accuracy by class. Finally, we highlight that our approach is fast. For a single image, the features can be extracted, augmented and classified within 0.4 seconds using a GPU implementation.

3.4 Conclusions

We demonstrated how filters from a CNN pretrained on natural images generalize to classifying 10 classes of non-dermoscopy skin images, outperforming previously published results. Our pipeline of using per-image-mean subtracted images, pooled-multi-scale feature extraction, and pooling across augmented feature space yielded consistent improvements to classification accuracy.

Chapter 4

Graph Geodesics to Find Progressively Similar Skin Lesion Images

4.1 Introduction

In the previous chapters, we discussed the importance of correctly diagnosing skin conditions. Many groups recognize the potential for computerized systems to analyze skin lesions and help reduce the burden on health care, and much work has gone into developing computerized systems to diagnose skin disorders [99], especially malignant melanoma, a common cancer which can be fatal [119]. Typically, such systems take as input a skin lesion image, and output either a discrete label or the probability that this lesion has a particular diagnosis. While knowing the probability that the image contains a particular type of skin lesion is a worthwhile goal, a disadvantage to this approach is that it is a "black box", where the user gains no insights into the automated diagnosis or of the underlying dataset of skin images.

4.1.1 Retrieving and Visualizing Similar Images

A different approach from classification that offers some insights into the dataset or diagnosis is to adopt an image retrieval based approach. For example, Celebi et al. [32] retrieved similar skin images using shape features extracted from segmented lesions and weighted these features to match the human perception of similar lesion shapes. Bunte et al. [29] extracted colour features from clinical skin images, learned a supervised transformation of these features, and retrieved images in a dataset based on the k nearest neighbours to these features. These returned images can be displayed to the user, giving insights into the appearance of similarly diseased images and allow the diagnosis to be inferred. Kawahara et al. [89] displayed a network graph visualization based on the nearest neighbours to a single query image, which allow users to efficiently search the space of similar lesion images.

Another approach to visualize *general images* was proposed by Hegde et al. [75], where rather than retrieving the k nearest images to a single query image, their approach uses two query images (a source and target) to retrieve a list of images that progress in visual similarity between the two images. They accomplish this by representing images as nodes in a graph, where the edges between nodes indicate their pair-wise distance, and the geodesic (shortest path) between source and target nodes represents a visually smooth progression of images. A similar approach for *general images* was recently implemented online [98], which is based on an experimental visualization tool that is part of Google Arts and Culture [97]. In other works, representing images as nodes to find an optimal path between nodes has been used to guide subject-template brain registration in MR images [81].

4.1.2 Potential Applications of Image Path Retrieval

In this work, we apply a similar method to find images of skin lesions that visually progress between a source and target lesion. This visualization approach may be useful for clinicians who wish to find reference images of hard to classify, visually challenging "borderline" cases across types of skin diseases (e.g., note the visually challenging aspects in distinguishing clark nevus from melanoma in Fig. 4.1 *bottom row*). Another use may be to show or predict the visual progression over time between a low-risk benign lesion to a malignant lesion (e.g., progression in Fig. 4.3 *bottom row*). This may give insights into how a disease may progress (e.g., Clark/Dysplastic nevi is potentially a precursor to melanoma and studies estimate that 20-30% of melanomas come from nevi [51]), or serve as a useful reference for patients to monitor and compare the progression of their own lesion. In these potential applications, the target images could be from either a set of predefined reference images, or the geodesics to each of the nearest unique diseases could be automatically shown.

4.1.3 Contributions

To the best of our knowledge, this is the first work that has applied geodesic paths to visualize skin lesion images. In contrast to previous work [75, 81, 97, 98], we propose to let each node in our graph represent images from two modalities (a dermoscopy and a clinical image), where the edge weights are influenced by both types of images. We apply an exponential function to the pair-wise dissimilarity measures, and show how this results in longer paths of higher quality without risking disconnected graphs. Finally, we propose measures to quantitatively evaluate the quality of our paths, which is lacking in prior work. These proposed quality measures are particularly important as without them, we would need to qualitatively inspect each path.

4.2 Methods

A skin lesion can be captured by both a dermatoscope (producing a *dermoscopy* image x_d), and a photo camera (producing a *clinical* image x_c), where the dermoscopy images show a more standardized view of the lesion, and the clinical images are non-standardized and often show additional contextual information (e.g., the body part the lesion is on) not available in the dermoscopy images. Given a dataset of skin lesions, the i-th skin lesion is represented by a pair of dermoscopy and clinical images $(x_d^{(i)}, x_c^{(i)})$. We create a graph where each pair of images $(x_d^{(i)}, x_c^{(i)})$ is represented by a single node $v^{(i)}$, and an edge $e^{(ij)}$ encodes the dissimilarity between nodes i and j. Our goal is to find a set of nodes $(v_0^{(s)}, v_1^{(i)}, \ldots, v_{R-1}^{(j)}, v_R^{(t)})$ of an unknown length R such that the initial node $v_0^{(s)}$ is a given source node (the superscript identifies the lesion, and the subscript indicates the position in the returned path), the R-th node is a given target node $v_R^{(t)}$, and the intermediate nodes $(v_1^{(i)}, \ldots, v_{R-1}^{(j)})$ represent lesions that visually progress between the source and target nodes. We find these intermediate nodes using Dijkstra's algorithm, which computes the geodesic between the source and target and returns a path of nodes representing a progression of visually similar lesions.

The key components that we now examine in detail are how to: extract image features that capture the salient properties of skin images, compute local dissimilarity between pairs of skin lesion images, weigh and connect the node edges using multi-modal images, and quantitatively evaluate the quality of the returned paths.

4.2.1 Skin Images as Deep Pretrained Neural Nets Responses

In Chapter 3, the responses of skin images with a CNN pretrained over ImageNet [140] were shown to be effective feature vectors for skin lesion classification despite the differences in appearance between skin lesions and natural images [90]. We use a similar approach to compute feature vectors as in [90], and for a particular image, extract the responses from the first fully connected layer of VGG16 [147], and average the responses over a set of predefined image augmentations,

$$\Phi(x)_m = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \phi(\pi(x - \mu))_m \tag{4.1}$$

where π is a function to augment an image (e.g., left-right flip); Π is the set of $|\Pi|$ number of image augmentations; $\phi(\cdot)_m$ extracts the m-th response of the first fully connected layer of VGG16; and, μ represent the mean pixel over the training data from ImageNet, which is subtracted from the skin lesion image x. The resulting feature vector $\Phi(x)$ represents a single lesion image by averaging the augmented responses over a single image, without increasing the dimensionality of the feature vector.

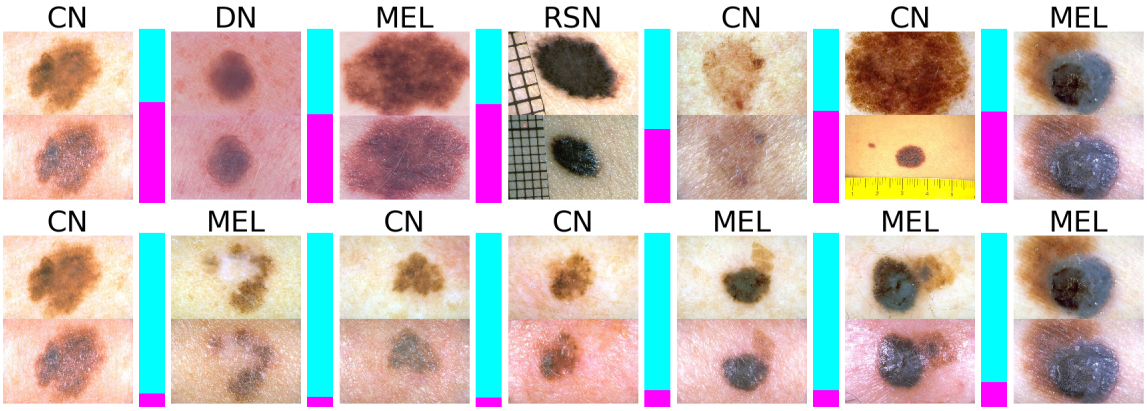


Figure 4.1: An example random path (*top*) and geodesic returned from the proposed method (*bottom*), where the *leftmost* and *rightmost* image represent the source and target nodes, respectively. The dermoscopy image is shown above the clinical image in each row. The *magenta bar* indicates the dissimilarity between adjacent images, where a higher bar indicates that they are more dissimilar.

4.2.2 Local Image Dissimilarity

Given two feature vectors $u, v \in \mathbb{R}^M$ (which represent the responses of two skin images), we compute the dissimilarity between them as the cosine distance raised to the p-th power,

$$\mathcal{D}(u, v) = \left(1 - \frac{\sum_i^M u_i v_i}{\sqrt{\sum_i^M u_i^2}\sqrt{\sum_i^M v_i^2}}\right)^p \tag{4.2}$$

where setting $p \neq 1$ non-linearly changes the dissimilarity between vectors. By using a high p (e.g., $p = 4$), we assign very low values to edges connecting similar images, thus encouraging geodesics to pass through nearby nodes of similar images, avoiding very short paths even in the case of complete graphs, i.e., fully connected graphs (further discussed in the Results section). Other distance measures are possible (e.g., L_1, L_2), and we found them to give empirically similar results. Fig. 4.1 shows the dissimilarity between pairs of images (dissimilarity is displayed in magenta using $p = 1$ for clarity).

4.2.3 Multi-modal Edge Weights

We define the edge weight $e^{(ij)}$ between nodes i and j as a weighted sum based on both the dermoscopy and clinical images,

$$e^{(ij)} = \alpha\mathcal{D}\left(\Phi(x_d^{(i)}), \Phi(x_d^{(j)})\right) + (1 - \alpha)\mathcal{D}\left(\Phi(x_c^{(i)}), \Phi(x_c^{(j)})\right) \tag{4.3}$$

where $\mathcal{D}(\cdot)$ is a function that computes the dissimilarity (Eq. 4.2) between the feature vectors $\Phi(\cdot)$ computed in Eq. 4.1; and α weighs the influence of the dermoscopy and clinical images ($0 \leq \alpha \leq 1$). Increasing α causes an edge to be more influenced by the dermoscopy

image than the clinical image, which may be desired as dermoscopy images contain more salient lesion properties.

4.2.4 Node Connectivity

To form the graph, we must decide on the connectivity of nodes. This can be done by connecting the k nearest neighbours (where nearest is defined via Eq. 4.2) to each node with an edge. However, choosing k is challenging as a large k (e.g., a complete graph) increases computational complexity and can lead to very short paths being returned when a direct edge exists between any pair of source and target nodes. Too small a k can lead to disconnected graphs, where no path exists between the source and target nodes. In the Results, we experiment with different values of k and show that by setting a high value of p in Eq. 4.3, the returned paths remain longer even in the case of complete graphs.

4.2.5 Surrogate Measures of Path Quality

While we provide qualitative results through visualizing the returned paths (Fig. 4.3), we also propose the following measures to quantitatively evaluate the quality of the returned paths. We define a *quality path* as a smooth visual progression of images. However, this definition is hard to precisely define and directly measure. Thus we propose a surrogate measure that uses the diagnoses of the lesions, as a skin lesion x is often accompanied with a corresponding clinical diagnosis y (e.g., melanoma, nevus) indicating the disease type, where y is an attribute of node v. Our assumption is that lesions with the same diagnosis will likely be visually similar, and that a high quality path will have a smooth progression with respect to the lesion diagnosis. In order to give a high cost to paths that frequently change neighbouring labels, we define the *transition cost* as,

$$\text{trans}(v_0, v_1, \ldots, v_{R-1}, v_R) = \frac{1}{R-1} \sum_{r=1}^{R} (1 - \delta(y_r, y_{r-1})) \tag{4.4}$$

where R is the number of nodes in the returned path; and y_r indicates the skin lesion diagnosis for the r-th returned path node corresponding to node v_r (e.g., y_0 and y_R correspond to the labels of the source and target nodes v_0, v_R, respectively). The Kronecker delta function $\delta(\cdot)$ returns 1 if the two labels have the same class, and 0 otherwise.

Our second surrogate quality measure quantifies the progression of the 7-point score between the source and target nodes. The 7-point score is a clinical measure of melanoma based on the visual presence of seven criteria (e.g., irregular streaks) within a lesion [9]. The weighted sum of these seven criteria form the 7-point score $\tau \in \mathbb{Z}$, where a higher score indicates the presence of more criteria associated with melanoma (further details in Section 6.2.4). We assume that a quality path will have 7-point scores that smoothly progress from a low to high score, as higher scores indicate the presence of lesion more indicative of

melanoma (and vice versa). We define the *progression cost* as,

$$\text{progress}(v_0, \ldots, v_R) = \frac{1}{R} \sum_{r=1}^{R} \left(\max \left[(\text{sgn}(\tau_0 - \tau_R)(\tau_r - \tau_{r-1})) \right], 0 \right]) \tag{4.5}$$

where the 7-point score τ is an attribute of node v (e.g., τ_0 and τ_R correspond to the 7-point score of the source v_0 and target v_R nodes, respectively); and, $\text{sgn}(\cdot)$ returns the sign of the difference between the source and target node scores,

$$\text{sgn}(a) = \begin{cases} 1, & \text{if } a = 0 \\ \frac{a}{|a|}, & \text{otherwise.} \end{cases} \tag{4.6}$$

This measure returns a cost of 0 if the 7-point score consistently decreases, increases, or remains constant along the path between the source and target nodes, and penalizes by the magnitude of the change otherwise. This approach, however, will always compute a 0 cost if the path only consists of the source and target nodes. As this is a degenerate case, we ignore the progression costs for paths of length two when computing results, and note that this measure is biased to return lower costs for shorter paths, and is thus most informative when comparing paths with the same number of nodes.

4.3 Results

4.3.1 Data

We test our proposed approach and surrogate measures using the Interactive Atlas of Dermoscopy [5] skin dataset. This dataset contains 1,011 cases of skin lesions, where all but four cases are captured by both a clinical x_c and dermoscopy x_d image (in the four cases missing x_c, we set $x_c = x_d$). Each case has a class label y that represents a known lesion diagnosis, and a 7-point score τ. The diagnosis y can take on one of the 15 class labels: basal cell carcinoma (BCC), blue nevus (BN), clark nevus (CN), combined nevus (CBN), congenital nevus (CGN), dermal nevus (DN), dermatofibroma (DF), lentigo (LT), melanoma (MEL), melanosis (MLS), miscellaneous (MISC), recurrent nevus (RN), reed or spitz nevus (RSN), seborrheic keratosis (SK), and vascular lesion (VL). The 7-point score $\tau \in \mathbb{Z}$ ranges between 0 and 7 (in this dataset), where a higher score indicates the lesion has visual properties more indicative of melanoma. The lesion diagnosis and the 7-point score are only used to quantify the quality of the returned paths, and are not used to form the graph. We randomly select a set of 1000 pairs of source and target nodes which are used across all experiments.

4.3.2 Recovering Synthetic Paths

We start by testing if our proposed approach can recover the path of images created by a progressive synthetic transformation. To do this, we crop the image by removing 15% of the

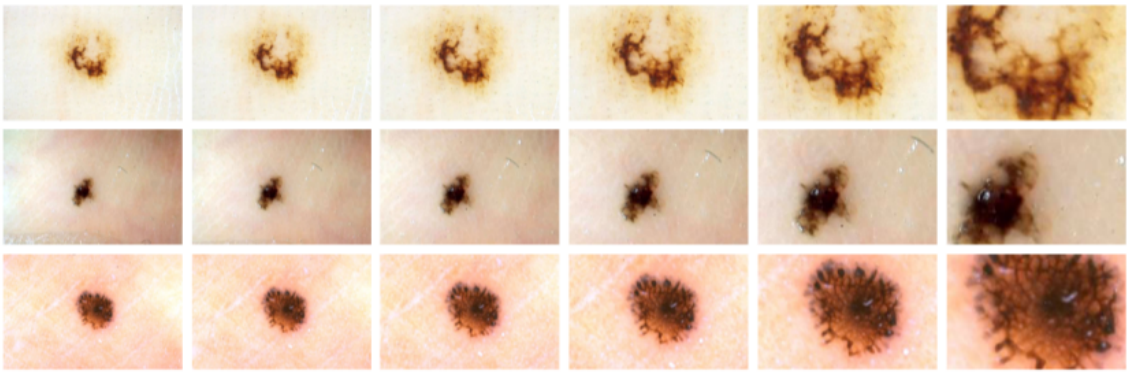


Figure 4.2: Synthetic examples: Here the *leftmost* images represent the source nodes, which belong to the original (non-enlarged) dermoscopy images in the dataset. The *rightmost* images represent target nodes, which were the last of the progressively enlarged images. The returned geodesic path is represented by the images in between. Note that the returned geodesic included all five synthetic images, in proper order of increasing enlargement.

pixels at the borders of the images, and repeat this five times. This progressively enlarges the lesion over a series of five images. We added these five synthetic images to our dataset, select the original image as the source and the final synthetic image as the target ($p = 4$ and $k = 30$). We find our approach not only recovers all synthetic images, but it recovers the correct sequence of synthetic images, i.e. in the order they were synthesized (Fig. 4.2), indicating that this approach and the feature vectors are sensitive to scale despite the CNN being trained on images at multiple scales.

4.3.3 Retrieving Paths from a Complete and Non-Complete Graphs

For our first experiment, in Table 4.1 **row 4.1.1** (*complete graph with* $p = 1$) we report results using a complete graph (i.e., $k = 1{,}011$) using only the dermoscopy images (i.e., $\alpha = 1$ in Eq. 4.1) and setting $p = 1$ in Eq. 4.3. We observe that when using a complete graph, the returned paths often consist of only the source and target nodes as their shared edge yields the shortest path. This experiment highlights the need to either prune the edges in the graph or modify the edge weights. Following the approach of [98], we form a new graph where each node is connected to its $k = 30$ neighbours. **Row 4.1.2** (*non-complete graph with* $p = 1$) shows that restricting the node connectivity increases the number of nodes in the returned path and improves the transition cost (note that the progression cost performs worse as it is biased towards paths with fewer nodes, and is thus most informative when comparing paths with a similar number of nodes).

4.3.4 Paths with Exponential Edge Weights

While decreasing node connectivity (i.e., lowering k) results in longer paths, care must be taken when choosing k, as reducing k increases the risk of forming disconnected graphs

Table 4.1: Quantitative results of the returned paths using the proposed surrogate quality measures. The *Img.* column indicates if the input was a dermoscopy image x_d, clinical image x_c, or included both. k represents the number of nearest neighbours used to form edges that connect nodes. *Aug.* indicates if the image was augmented or not when forming the image feature vector. *Trans., Progress.,* indicates the average and standard deviation transition and progression cost as defined in the text. *Num. Path* shows the average and standard deviation number of nodes in the computed path.

Exp.	Img.	Aug.	p	k	Ordered	Trans.	Progress.*	Num. Path
4.1.1	x_d	✗	1	1011	min-path	0.76 ± 0.42	0.13 ± 0.41	2.02 ± 0.13
4.1.2 [98]	x_d	✗	1	30	min-path	0.64 ± 0.34	0.42 ± 0.44	3.59 ± 0.85
4.1.3	x_d	✗	4	1011	min-path	0.56 ± 0.26	0.63 ± 0.33	8.11 ± 2.87
4.1.4	x_d	✗	4	30	min-path	0.56 ± 0.26	0.63 ± 0.33	8.12 ± 2.87
4.1.5	x_d	✓	4	30	min-path	0.55 ± 0.25	0.59 ± 0.29	9.16 ± 3.62
4.1.6	-	-	-	-	random	0.77 ± 0.19	0.92 ± 0.40	9.16 ± 3.62
4.1.7	x_d	✓	-	-	linear	0.58 ± 0.25	0.74 ± 0.36	9.16 ± 3.62
4.1.8	x_c	✗	4	30	min-path	0.65 ± 0.18	0.71 ± 0.34	10.64 ± 5.08
4.1.9	x_d, x_c	✗	4	30	min-path	0.45 ± 0.24	0.56 ± 0.32	7.90 ± 3.27
4.1.10	x_d, x_c	✓	4	30	min-path	0.45 ± 0.23	0.60 ± 0.29	8.86 ± 3.73

*The *Progress.* column differs from [92] due to an earlier mistake in computing the seven-point scores.

where no path exists between a source and target node. Thus instead of pruning edges, our next experiment (**row 4.1.3** *complete graph with* $p = 4$) shows how applying an exponential function (i.e., $p = 4$ in Eq. 4.3) to the dissimilarity function results in longer paths of higher quality even in a complete graph. By removing the need to prune graphs (i.e., choose k), we guarantee the existence of a path, while still preventing short paths. If we are not concerned with disconnected graphs, we can combine edge pruning using k neighbours with the increased p, to match the computational efficiency of a pruned graph without any quality degradation (**row 4.1.4** *non-complete graph with* $p = 4$). For the remaining experiments, we use $p = 4$ and non-complete graphs with $k = 30$, as our graphs remained connected.

4.3.5 Comparing Random and Linearly Interpolated Path

In **row 4.1.5** (*augmented images*) we augment the feature vector with left-right image flips (Eq. 4.1), which results in longer geodesics paths and minor improvements to the path quality. We form a path with an equal number of nodes as those returned in the geodesic path in the previous experiment (from row 4.1.5) by randomly sampling nodes (without replacement). As the labels in our dataset are highly imbalanced, these random paths give us a baseline quality score (**row 4.1.6** *random paths*). We also compare our method by ignoring the graph, and instead using linearly interpolated feature vectors between the source and target feature vectors. These interpolated feature vectors are uniformly separated to match the number of returned nodes in row 4.1.5. The nearest unique neighbour to this

interpolated feature vector is used to form the path. **Row 4.1.7** (*linear paths*) shows that this approach yields paths of worse quality when compared to using graph geodesics. We highlight that the graph geodesic approach has the additional advantage of automatically determining the number of nodes in the path, whereas the linearly interpolated approach requires this to be specified (we set it equal to the length of the geodesic path).

4.3.6 Using Clinical Image Features

In **row 4.1.8** (*clinical images*) we use only the clinical image (i.e., $\alpha = 0$ in Eq. 4.3) and notice a marked decrease in the quality of the paths when compared to dermoscopy images. This is expected since dermoscopy images are more standardized and focused on the lesion, while clinical images have a non-standard field of view and can capture background artifacts.

4.3.7 Combining Dermoscopy and Clinical Image Features

In **row 4.1.9** (*dermoscopy and clinical images*) we include both the clinical and dermoscopy images, weighting the dermoscopy images higher (i.e., $\alpha = 0.8$ in Eq. 4.3) as the dermoscopy images better capture the salient lesion features and avoid irrelevant background artifacts. The returned paths now respect both imaging modalities, yielding improvements to the quality of the paths, most noticeable with transition costs. Finally, in **row 4.1.10** (*full approach*) we show the full proposed approach, which uses augmented images from both modalities with the dissimilarity measure raised to the power of $p = 4$ on a non-complete graph. While the path quality measures remain similar to the previous experiment, the total path length increases.

4.4 Conclusions

We proposed a method to visualize a smooth progression of similar skin lesion images between two skin lesions. Our graph geodesic based approach applies an exponential dissimilarity function and considers information from multiple modalities (clinical and dermoscopy images) to form the graph edges, leading to longer paths of higher quality. We proposed surrogate measures of path quality based on the diagnostic labels of the skin lesions to quantitatively assess the resulting paths. Future work would explore how to improve the feature vectors that represent the skin images (e.g., fine-tuning the CNN over a skin dataset), and examine how to make the progression quality measure less sensitive to the length of the path.

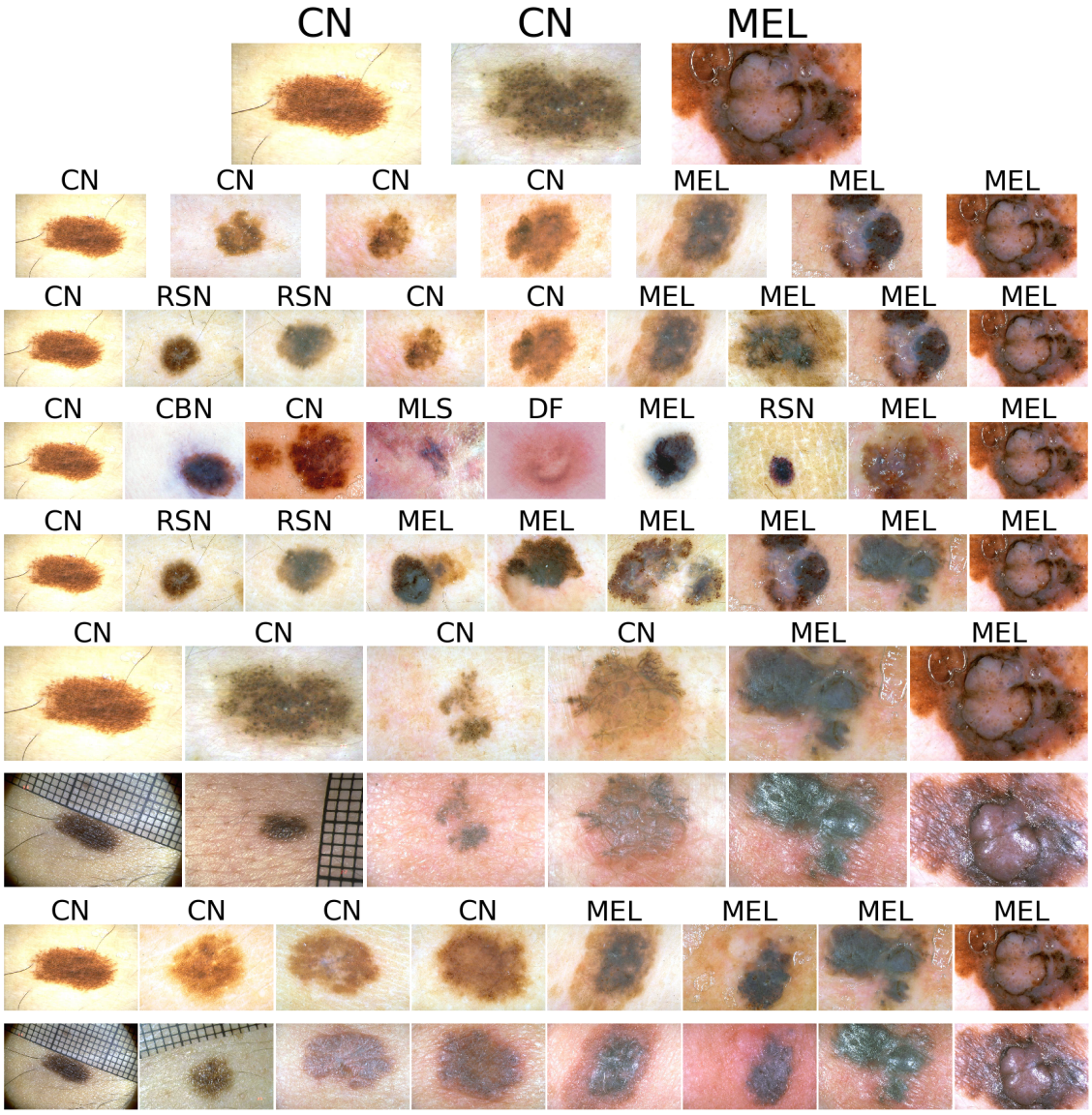


Figure 4.3: Visualizing Paths. The *leftmost* and *rightmost* dermoscopy images are the given source (clark nevus) and target (melanoma) node, where the images in each row in between them correspond to the computed geodesic/minimal path. Each row, starting from the *top* to *bottom* row, correspond to the following experiments in Table 4.1: 4.1.2 (*non-complete graph with* $p = 1$), 4.1.4 (*non-complete graph with* $p = 4$), 4.1.5 (*augmented images*), 4.1.6 (*random paths*), 4.1.7 (*linear paths*), 4.1.9 (*dermoscopy and clinical images*), and 4.1.10 (*full approach*). The geodesic of Experiments 4.1.9 and 4.1.10 incorporates clinical images, shown directly below the dermoscopy images.

Chapter 5

Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers

5.1 Introduction

The World Health Organization estimates that globally each year, between two and three million nonmelanoma skin cancers are diagnosed, and 130,000 melanoma skin cancers occur [170]. Diagnosing different types of skin lesions is needed to determine appropriate treatment, and computerized systems that classify skin lesions from skin images may serve as an important screening or second opinion tool. While considerable research has focused on computerized diagnosis of melanoma skin lesions [99], less work has focused on the more common nonmelanoma skin cancers and on the general multi-class classification of skin lesions. In this chapter, we focus on predicting multiple types of skin lesions that includes both melanoma and nonmelanoma types of cancers.

5.1.1 Multi-resolution CNNs

CNNs generally learn based on an image of a single fixed resolution (e.g., Krizhevsky et al. [100]). However, this single resolution may not be optimal and depends on the scale of the objects within the image. Information from multiple image resolutions may be critical in capturing fine details, especially in the domain of medical images (e.g., to discriminate pathology). As such, other works have proposed different multi-scale approaches.

During testing, Sermanet et al. [144] used a fully convolutional neural network to extract predictions over multiple image resolutions and spatial locations and aggregated the predictions using a spatial max and averaging of scales. This simple aggregation approach, however, does not *learn* interactions across different resolutions (i.e., multi-resolution only applied during testing, not training).

He et al. [73] proposed a spatial pyramid pooling layer applied after the last convolutional layer to produce fixed-sized responses regardless of the image size. The CNN is trained on images of multiple resolutions *sequentially*, causing the CNN to learn parameters that generalize across image resolutions. However, each prediction is based only on a single input resolution, and interactions across multiple input image resolutions are not considered.

Bao et al. [14] proposed a multi-scale CNN trained and tested on image patches of different sizes (i.e., *different* field-of-view) simultaneously for segmentation, but did not explore multi-resolution input (i.e., *same* field-of-view at different resolutions) for whole image classification. Kamnitsan et al. [84] proposed a multi-scale dual-path 3D CNN for brain segmentation that, like the prior approach, considers *different* field-of-views.

In Chapter 3, to perform 10-class skin lesion classification of nonmelanoma and melanoma skin lesions, Kawahara et al. [85] performed a two step process: first, using a CNN pretrained over ImageNet [140], they extracted image features at two different image resolutions, and second, these features were concatenated and used to train a linear classifier. While the approach in Chapter 3 does consider multiple image resolutions, only a final linear classifier learns interactions across different image resolutions, and the CNN itself does not learn based on the input images.

5.1.2 Contributions

In this chapter, we propose a CNN for skin lesion classification that *learns interactions across multiple image resolutions* of the same image simultaneously through multiple network tracts. Unlike prior multi-scale architectures [14, 84], our network keeps the *same* field-of-view for image classification, uses auxiliary loss functions, and leverages parameters from existing pretrained CNNs. Leveraging pretrained CNN parameters (i.e., transfer learning) is especially useful with limited training images, and has resulted in consistent improvements in other medical image analysis tasks when compared to starting from random initialization [146]. Thus a key contribution of our work is to extend pretrained CNNs for multiple image resolutions, optimized end-to-end with a single objective function. We demonstrate that our proposed multi-tract CNN outperforms competing approaches over a public skin dataset.

5.2 Methods

We design a CNN to predict the true lesion class label y, given a skin lesion image x. Our CNN is composed of multiple tracts where each tract considers the same image at a different resolution using the same field-of-view. An end layer combines the responses from multiple resolutions into a single layer. We attach a *supervised loss layer* (i.e., layer with a loss function that compares predicted with true class labels) to these combined responses, thus making the final prediction a learned function of multiple resolutions of

the same image. This loss is backpropagated through all tracts causing the entire network to be optimized with respect to multiple image resolutions. We add auxiliary supervised loss layers to each tract, motivated by the work of Lee et al. [104], who found that adding additional "companion"/"auxiliary" supervised layers regularize the responses learned. In this work, auxiliary losses cause each tract to learn parameters that classify well at that particular resolution. At test time, we ignore the auxiliary classifiers and only use the final end classifier.

5.2.1 Converting a Pretrained CNN to Multi-Tract CNN

In order to train large CNNs with a limited skin dataset, we use a hybrid of the pretrained AlexNet [50, 100] architecture and parameters θ_p, (omitting the 1000-d ImageNet-specific output layer) for early network layers, and additional untrained layers for later network layers that learn only from skin images. To pass images of different resolutions through all the layers pretrained on a single resolution, we convert (keeping the trained parameters) fully-connected layers to convolutional layers, as convolutional layers allow for variable sized inputs [144].

For practical considerations (e.g., limited GPU memory), we limit our discussion and experiments to two tracts, although this approach is applicable to additional tracts/resolutions. We refer to the two tracts as the *upper tract*, which takes in a low-resolution image, and the *lower tract*, which takes in a high-resolution image. Our full proposed network is shown in Fig. 5.1.

We pass an image x_1, of the same image resolution that the pretrained network (AlexNet) was trained on to the *upper tract* of our network. This produces responses of size $1\times1\times4096$. We add an additional convolutional layer with untrained (i.e., randomly initialized) parameters $\theta_t^{(1)}$, which produces responses of lower dimensionality $1\times1\times256$. To the *lower tract*, we pass x_2, which is the same image but with a resolution greater than that of x_1. After being convolved with the pretrained parameters, the lower tract produces responses of $m\times m\times4096$ (Fig. 5.1 *lower green box*). Other works have reduced this $m\times m$ dimensionality through pooling [73, 85, 144], but in this work, we add additional untrained $1\times1\times4096$ and $m \times m \times 64$ convolutional filters, $\theta_t^{(2)}$, that *learn* to reduce the dimensionality to $1\times1\times256$. By using these two layers instead of a single fully-connected layer, we significantly reduce the amount of needed parameters. Auxiliary supervised loss layers with untrained parameters, $\theta_l^{(1)}, \theta_l^{(2)}$, are added to the upper and lower tract responses.

An untrained convolutional layer takes as input the 256-dimensional responses from both tracts. We add a supervised loss layer to these combined responses making the final prediction a function of the image taken at two different resolutions. In order to reduce the total number of independent parameters in our model, we are inspired by the work on Siamese nets [38] to share the AlexNet weights θ_p, across the upper and lower tracts. This means that updates to θ_p will be based on both image resolutions. Finally, rather than

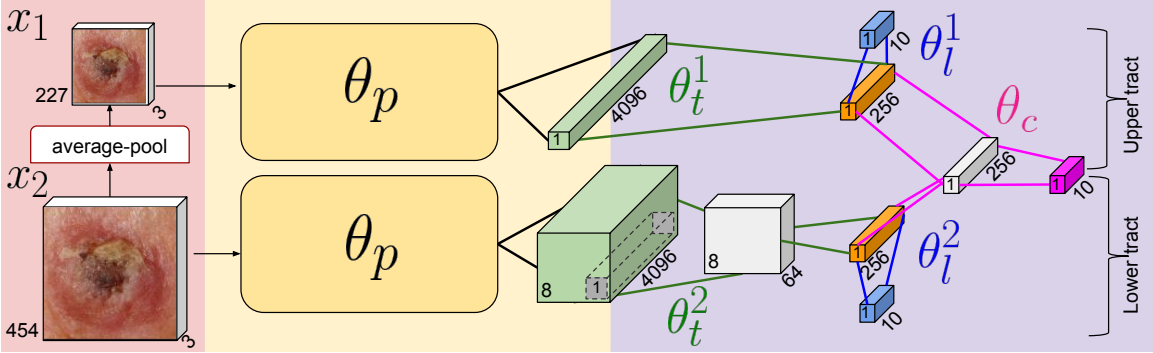


Figure 5.1: The proposed two-tract fully convolutional multi-resolution neural network. The highest resolution image x_2 is averaged-pooled to create a low-resolution image x_1, which serves as input to the *upper tract*. x_2 is fed to the *lower tract* to extract responses at a finer scale. As all layers are convolutional layers, a larger input produces larger responses (*green lower box*). After the layers with pretrained parameters θ_p, additional layers with unshared trainable parameters θ_t, are added. Each tract has a supervised auxiliary loss layer (*blue box*). The responses from both image resolutions are combined and an output layer makes the final prediction (*pink box*). Spatial dimensions (e.g., 8 mean 8×8) are given inside each box, and the number of channels are shown alongside each box.

storing separate image resolutions of the same image, we only store the highest desired resolution. Within the network architecture itself, we average-pool the high-resolution image to the desired low-resolution scale, allowing for more efficient storage.

5.2.2 Multi-Tract Loss and Optimization

Our network has a supervised data loss term that considers the combined high and low resolution images, as well as auxiliary data losses that each only considers the responses from a single image resolution. These equally weighted losses are averaged over a mini-batch of N training instances along with a regularization over the parameters,

$$\mathcal{L}(x, y, \theta) = \frac{\lambda}{N} \sum_{i=1}^{N} \left(\ell(x^{(i)}, y^{(i)}; \theta_p, \theta_t, \theta_c) + \sum_{j=1}^{N_{\text{aux}}} \ell(x_j^{(i)}, y^{(i)}; \theta_p, \theta_t^{(j)}, \theta_l^{(j)}) \right) + \gamma ||\theta|| \qquad (5.1)$$

where $\ell(\cdot)$ is the cross-entropy loss using a softmax activation function; the i-th image is transformed into the j-th resolution $x_j^{(i)}$; and, $y^{(i)}$ is the ground truth class label of $x^{(i)}$. The parameters $\theta = \{\theta_p, \theta_t, \theta_c, \theta_l\}$ are composed of the shared pretrained parameters θ_p, the unshared tract parameters $\theta_t = \{\theta_t^{(j)}\}$ where j indicates the tract, the parameters connecting the j-th auxiliary loss $\theta_l = \{\theta_l^{(j)}\}$, and the parameters connecting the tracts together θ_c. $||\theta||$ is the L2 regularization over the parameters. λ, γ weight the terms, where $\lambda = \frac{1}{N_{\text{aux}}+1}$ and N_{aux} is the number of auxiliary supervised layers in the network (e.g., $N_{\text{aux}} = 2$).

We update our network parameters θ using stochastic gradient descent with mini-batches. Thus, for the k+1 iteration, we compute $\theta_{(k+1)}$, from the previous k-th iteration

parameters $\theta_{(k)}$ and parameter updates $U_{(k)}$ as,

$$U_{(k+1)} = \mu U_{(k)} - \alpha \nabla \mathcal{L}(\theta_{(k)}) \qquad \text{and} \qquad \theta_{(k+1)} = \theta_{(k)} + U_{(k+1)}, \tag{5.2}$$

using a low learning rate $\alpha = 10^{-4}$, as much of the CNN is pretrained; $\nabla \mathcal{L}(\theta_{(k)})$ are the gradients of Eq. 5.1; and μ is a momentum parameter. We use Caffe [82] to implement our architecture and optimize Eq. 5.1 with mini-batches of size $N = 15$ (lowers GPU memory to allow for multiple tracts). As common in the literature [100], we set $\mu = 0.9$ and $\gamma = 0.0005$.

5.3 Results

We used the Dermofit Image Library[1] to test our proposed method. This dataset contains 1,300 skin lesion images from 10 classes: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Melanocytic Nevus/Mole (ML), Squamous Cell Carcinoma (SCC), Seborrhoeic Keratosis (SK), Intraepithelial Carcinoma (IEC), Pyogenic Granuloma (PYO), Haemangioma (VSC), Dermatofibroma (DF), and Malignant Melanoma (MEL). We randomly divided the dataset into three subsets of approximately the same number of images and class distributions. One subset is used to train (i.e., optimize Eq. 5.1), validate (e.g., test design decisions), and test. We resized the x_1 image to 227×227 and x_2 to 454×454. Each image was normalized by subtracting the per-image-mean intensity as in [85].

5.3.1 Performance with Fixed Pretrained Parameters

For our first experiments (Table 5.1 rows *a-c*), we implemented the two-step approach of Kawahara et al. [85] described in Chapter 3, by extracting responses from the sixth layer (FC6) of the pretrained AlexNet for images x_1 and x_2, and max-pooled the spatial responses of x_2. As in Chapter 3 [85], these extracted responses are used to train a logistic regression classifier. We report the accuracy for classifying x_1 and x_2 individually, and on the concatenated responses from the two image resolutions (note this experimental setup only uses half of the training images that Chapter 3 [85] did).

5.3.2 Single Resolution with Trainable Parameters

Our next experiments (Table 5.1 rows *d,e*) show that our hybrid use of pretrained and additional skin-lesion trained layers improved classification accuracy. We split the two-tract network into upper and lower tracts and train each separately on a single resolution. For a fair comparison, we doubled the number of nodes in the layer before the auxiliary loss layer (i.e., Fig. 5.1 *orange layer*) to closely match the number of independent parameters within the two-tract model. The accuracy of the one-tract single-resolution model (rows

[1] `https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html`

Table 5.1: Experimental results. *image res.* shows the image resolution in the train/test phase (e.g., 227/454 means image size 227×227 and 454×454). We report the classification accuracy for the *valid* and *test* datasets. Rows *a-i* use multi-resolution versions of an image spanning the same field-of-view. Rows *j,k* use augmented image views, where row *k* combines the multi-resolution approach with augmented views.

	method	image res.	valid	test	
(a)	FC6+LogReg	227	0.674	0.705	single view
(b)	FC6+LogReg	454	0.649	0.700	
(c)	Kawahara et al. 2016 [85]*	227/454	0.684	0.741	
(d)	1-tract (*ours*)	227	0.733	0.741	
(e)	1-tract (*ours*)	454	0.737	0.759	
(f)	1-tract + Sermanet et al. 2014 [144]	227/454 (test)	0.719	0.748	
(g)	He et al. 2014 [73] (SPP)	224/448	0.688	0.711	
(h)	2-tract 0-aux-losses (*ours*)	227/454	0.723	0.755	
(i)	2-tract 2-aux-losses (*ours*)	227/454	**0.751**	**0.773**	
(j)	1-tract (*ours*)	454	0.760	0.775	aug. view
(k)	2-tract 2-aux-losses (*ours*)	227/454	**0.781**	**0.795**	

*Results differ from Chapter 3 [85] due to differences in experimental setup.

d,e) improved over rows *a,b*, but is less than our proposed model (row *i*), indicating that considering multiple resolutions within our two-tract architecture improves accuracy.

5.3.3 Comparing Other Multi-resolution CNNs

Table 5.1 row *f* details the results of applying the classification approach of Sermanet et al. [144] to aggregate the CNN responses from multiple image resolutions. To implement their classification approach, we pass high-resolution x_2 images through the one-tract model (row *d*) trained on low-resolution x_1 images to produce class responses with spatial dimensions. We take the maximum spatial response and average it with the class responses computed from the low-resolution image to compute a class unnormalized probability vector.

Table 5.1 row *g* uses He et al. [73]'s Spatial Pyramid Pooling (SPP) approach, which learns CNN parameters from multiple image resolutions. To implement, we use the pre-trained Zeiler-Fergus (ZF) SPP network He et al. [73] provided (similar architecture to AlexNet) and replace their final output layer with our own. We train over ≈11 epochs before switching between 224×224 and 448×448 image resolutions, repeating 20 times for 9000 iterations (more iterations did not improve results). Each image resolution is fine-tuned for 1000 iterations. During testing, we averaged the CNN's output class responses from both resolutions.

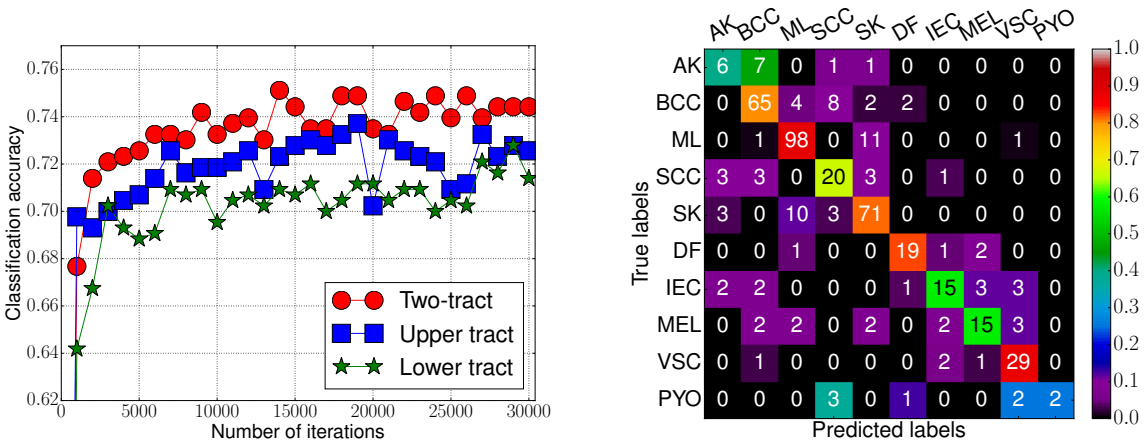


Figure 5.2: (*left*) A comparison of the classification accuracy of the individual upper and lower single resolution-tracts with the two-tracts. Integrating multiple image resolutions yields higher accuracy. By using pretrained parameters, we reach a high accuracy within a short number of iterations. (*right*) The confusion matrix over the 10-classes from our test data using our proposed multi-tract CNN (heatmap indicates class-specific classification accuracy normalized across rows).

5.3.4 Multi-tract with Auxiliary Losses

In Table 5.1 rows *h,i* we show results using our two-tract multi-resolution architecture. Without auxiliary losses (row *h*), the two-tract model performs worse than the single tract (row *e*), highlighting the need to include the auxiliary loss functions (Eq. 5.1) to achieve the highest accuracy (row *i*). Note that we outperform [85], which was shown to outperform [12, 105], and that [12, 105] were non-CNN based approaches specifically designed for this dataset. The confusion matrix over the test data is shown in Fig. 5.2 (*right*). We ran additional experiments to cross-validate over the two other folds and obtained a statistically significant difference between the baseline of [85] (using the approach from row *c*) and our two-tract approach (row *i*) with a mid-p McNemar's test, p=0.0155.

We compare the accuracy of the final output classifier with the accuracy of the auxiliary classifiers (Fig. 5.2 (*left*)). Generally, the final classifier has a higher accuracy, indicating that this classifier (which considers the same responses as each auxiliary classifier) has learnt to combine responses from multiple image resolutions, and that this improves classification accuracy. This plot also highlights the advantage of pretrained parameters, as high accuracy occurs within 5000 iterations (1 hour of training), using a low number (430) of training images.

In order to focus on the effects of our proposed architecture and multi-resolution input, the experiments in Table 5.1 rows *a-i* did not use data augmentation. Our final set of experiments demonstrates that our multi-resolution approach is *complementary* to the commonly used approach of training using different image views. We augment the training

images with left-right flips, and rotations. Row k combines augmented image views with multi-resolution input, resulting in further accuracy improvements when compared to using only augmented views (row j) and using only multiple resolutions (row i), highlighting that the proposed multi-resolution input complements existing image augmentation approaches.

We did not compare to [14, 84] as their approach was designed for 3D segmentation, and while their approach of taking as input different amount of spatial context is well motivated for patch-based segmentation of 3D volumes, it is less applicable to whole image classification. Further contributions we make that differ with their work include: pretrained CNNs for multiple resolutions, the use of auxiliary losses, and multi-resolution input.

Finally, we discuss possible reasons why successful approaches used in *computer vision datasets* (e.g., ImageNet [140] where images are captured at widely different scales), were found less effective for our skin diagnosis application (where dermatology images are captured at a similar scale). When the scale of objects widely differs, the SPP approach [73] to learn parameters that *generalize* over multiple scales, and the approach to *aggregate* responses over different scales [144], are desirable. However, in our case, where the objects' scale are roughly fixed, the different CNN-tracts learn to respond to characteristics that are *specific* to that resolution. This highlights how our proposed architecture is well designed for skin images captured at relatively fixed scales.

5.4 Conclusions

We presented a novel multi-tract CNN that extends pretrained CNNs for multi-resolution skin lesion classification using a hybrid of pretrained and skin-lesion trained parameters. Our approach captures interactions across multiple image resolutions simultaneously in a fully learned end-to-end optimization, and outperforms related competing approaches over a public skin lesion dataset.

Chapter 6

Seven-Point Checklist and Skin Lesion Classification using Multitask Multimodal Neural Nets

6.1 Introduction

Skin cancer is the most common malignancy in fair-skinned populations, and the incidences of melanoma and non-melanoma skin cancers are rising, resulting in high economic costs [4]. Early melanoma diagnosis appears to improve patient outcomes [19], and skin cancer detection can be improved through approaches such as screening patients with focused skin symptoms using physician-directed total body skin examinations [8].

Epiluminescence microscopy or dermoscopy, which is a noninvasive in-vivo imaging technique, uncovers detailed morphological and visual properties of pigmented lesions. Kittler et al. [96] reported that, for experienced dermatologists, the accuracy in diagnosing pigmented skin lesions improves when using dermoscopy compared to the unaided eye. However, accurate diagnosis is challenging for non-experts.

Pattern analysis, which subjectively assesses multiple subtle lesion features, is commonly used by experienced dermatologists to distinguish between benign and malignant skin tumours. To simplify diagnoses, rule-based diagnostic algorithms such as the ABCD rule [129] and the 7-point checklist [9] have been proposed and are commonly accepted [24]. In this work we focus on the 7-point checklist, which requires identifying seven dermoscopic criteria (Table 6.2) associated with melanoma, where each criteria is assigned a score. The lesion is diagnosed as melanoma when the sum of the scores exceeds a given threshold [7, 9]. Although some literature recommends pattern analysis over the 7-point checklist [31], some works report a trade-off between melanoma sensitivity and specificity. For example, among dermatology residents, the 7-point checklist was found to give higher sensitivity, but lower specificity than pattern analysis [31]. A similar result was found among experienced dermatologists using a lowered 7-point checklist threshold [7]. This indicates limitations with both approaches, and motivates additional study. Further, although the 7-point checklist

and pattern analysis diagnostic procedures are different, the 7-point checklist criteria are based on the criteria used in the process of pattern analysis [6]. Detecting these criteria may aid with more interpretable diagnostic models regardless of the preferred diagnostic procedure (e.g., report the presence of dermoscopic criteria associated with malignancy, retrieve images with specific criteria).

Computer aided approaches to classify dermoscopy images have attracted significant research attention as automated analysis has the potential to empower patients with timely, reproducible diagnoses, especially in remote communities with limited clinical access. Furthermore, the increasing prevalence of mobile and relatively inexpensive dermatoscopes, suggests increased access to personal dermoscopy imaging devices.

6.1.1 Approaches to Detect the 7-point Checklist Criteria

Many previous works focus on detecting a single criterion from the 7-point checklist. For example, Mirzaalian et al. [125] detected absent, regular, and irregular streaks by enhancing streaks using Hessian based tubular filters. They tested on 99 dermoscopy images from the Argenziano et al. [5] dataset. Madooie et al. [116] detected the presence of blue white veils by mapping image regions to a discrete set of Munsell colours, using 223 images also selected from [5].

A few works detected the entire 7-point checklist [17, 55, 165]. Fabbrocini et al. [55] detected all 7-point checklist criteria by designing separate pipelines that consider each criterion's unique characteristics. However, each pipeline adds complexity and requires careful tuning of hyper-parameters. For example, to detect irregular streaks, precise lesion border detection is required to compute an "irregularity" index, which considers how the lesion border differs from a straight line when the lesion is divided into segments. To detect irregular dots and globules, they applied statistical region merging [131] to find candidate dark segments, extracted morphological features, and applied thresholds (set experimentally) to detect rounded areas. Similar customized pipelines were set for all criteria. Wadhawan et al. [165] also proposed a system to detect all 7-point checklist criteria. Taking a machine learning approach, they extracted human engineered features (e.g., Haar wavelet, local binary patterns, colour histograms) from a segmented region of interest. For each criterion, they selected a subset of features that correlated well with the criterion, and used these subsets to train a support vector machine. For evaluation, they considered 385 low difficulty images from the Argenziano et al. [5] dataset, out of which 347 could be segmented to create satisfactory lesion boundaries.

6.1.2 Approaches to Directly Classify Skin Conditions

Rather than detecting the 7-point checklist to infer a melanoma diagnosis, other works have explored directly classifying the disease from the image. For instance, the International Skin Imaging Collaboration's skin lesion classification challenge [66], asks participants to directly

classify benign from malignant lesions. The top performing classification approach by Yu et al. [176] fine-tuned a Residual Neural Network [74] pretrained over ImageNet [140]. Over the DermoFit dataset [12], which is composed of 10 types of skin conditions in standardized clinical images, in Chapter 3, Kawahara et al. [85] demonstrated how using features from a neural network pretrained over ImageNet to classify skin diseases outperformed approaches that rely on handcrafted human engineered features [12, 105]. Over the Argenziano et al. [5] dataset, Menegola et al. [122] showed that fine-tuning a neural network pretrained only over ImageNet [140], performed better than training a neural network from scratch, or when pretrained over a dataset that included retinopathy images.

6.1.3 Contributions

This is the first work that predicts the entire 7-point criteria and the diagnosis (including the melanoma classification) in a single optimization, where predictions are derived from a multi-modal CNN that considers clinical images, dermoscopy images, and meta-data. Further, we show how our proposed deep architecture is used for three common tasks: classification, extracting feature vectors for image retrieval that consider clinical criteria, and localization of discriminative regions. We also publicly release the Argenziano et al. [5] dataset. While this dataset has been partly used in other publications (e.g., [116, 122, 141, 165]), it has not been readily available to the public. This dataset has been noted to have "excellent interobserver agreements" [83], and was used to teach dermatologists [31, 110], suggesting that it is a suitable source for training machine learning algorithms.

6.2 Methods

Given a dataset of skin lesions, we define each unique lesion as a *case*. The i-th case can have multiple types of information associated with it, such as a dermoscopy $x_d^{(i)}$ image, a clinical $x_c^{(i)}$ image, and patient meta-data $x_m^{(i)}$. Dermoscopy images x_d are captured with a dermatoscope and offer a standardized field of view and controlled acquisition (e.g., lighting and field of view). Clinical images x_c are less standardized, taken at various fields of view, and can contain image artefacts (e.g., a ruler to measure the lesion). Patient meta-data x_m includes other types of information, such as patient gender and lesion location.

Each case has a set of *categories* associated with it, assigned by a dermatologist. The categories are composed of labels for a diagnosis and the 7-point checklist criteria. The *diagnosis* $y^{(i)}$ assigns an overall skin condition label to the i-th case (e.g., melanoma, basal cell carcinoma). The *7-point checklist criteria* $z^{(i)}$ consists of seven criteria that identify skin lesion properties that are indicative of melanoma [9], where the j-th criteria in the seven-point checklist $z_j^{(i)}$ has different labels associated with it. For example, "pigment network" is one of the 7-point checklist criteria with three labels: atypical, typical, and absent. We use the term *categories* to refer to both the diagnosis and the 7-point checklist criteria,

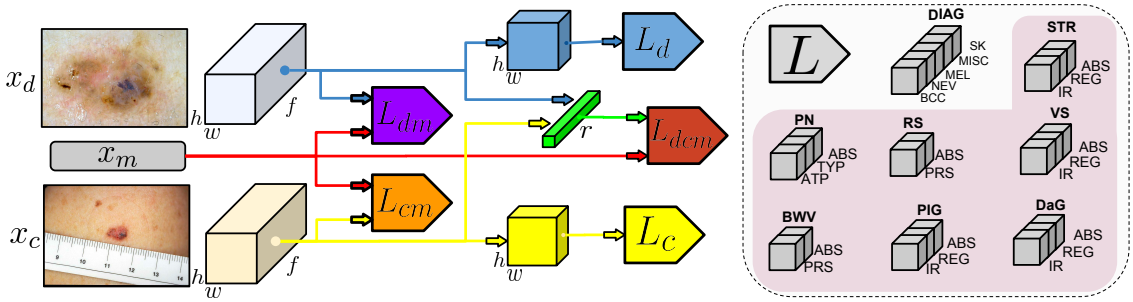


Figure 6.1: The proposed architecture considers dermoscopy images x_d, clinical images x_c, and meta-data x_m when classifying all 7-point criteria and diagnosis. Each multi-task loss (L *block*) is trained on different combinations of the input modalities (e.g., L_{dm} is a function of x_d and x_m). As each L block gives predictions based on the data it was trained on, this single model is robust to missing data at inference time. The *blue* and *yellow* blocks immediately before the multi-task loss indicates the layer that is used to localize the discriminative regions. The *green bar* indicates the multi-modal feature vector used for image retrieval.

while the term *labels* refers to items within a specific category. Each category has a variable number N_c of possible labels c (e.g., the "pigment network" category has $N_c = 3$ labels). The full list of categories and labels are given in Table 6.1 and Table 6.2.

6.2.1 Multi-Modal Multi-Task Loss Function

Rather than developing a separate model or pipeline for each individual category as is commonly done, we present a single model that predicts all labels within each category in a single optimization. We use a CNN, which consists of a designed architecture and a set of trainable parameters θ. Given the case data x (which represents different combinations of the input modalities), we define a *multi-task* loss function L for all eight (7-point checklist and a diagnosis) categories as,

$$L(x, y, z; \theta) = \ell(x, y; \theta) + \sum_{j=1}^{7} \ell(x, z_j; \theta) \tag{6.1}$$

where $\ell(\cdot)$ is a weighted categorical cross-entropy loss defined as,

$$\ell(x, c; \theta) = -\frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{j=1}^{N_c} w(c)_j \cdot c_j^{(i)} \cdot \log\left(\phi(x^{(i)}; \theta)_{c,j}\right) \tag{6.2}$$

where $c \in \mathbb{Z}^{N_b \times N_c}$ is the 1-hot-encoded ground-truth labels for a single category; N_b is the number of cases in a mini-batch; and, N_c is the number of labels for the c-th category. The weighting function $w(c)_j$ gives a higher weight to infrequent labels and is defined in Eq. 6.5. $\phi(x^{(i)}; \theta)_{c,j}$ is the probability predicted by the neural network parameterised by θ for the j-th label of the c-th category given input $x^{(i)}$. The multi-task loss (Eq. 6.1) is a function

of the input modalities x; however, the available data x may vary by case (e.g., missing meta-data). In order to handle these cases, we define a *multi-modal multi-task* loss function that considers multiple combinations of the input modalities as,

$$\begin{aligned}\mathcal{L}(x_d, x_c, x_m, y, z; \theta) = L((x_d, x_c, x_m), y, z; \theta_{dcm}) \\ + L(x_d, y, z; \theta_d) + L((x_d, x_m), y, z; \theta_{dm}) \\ + L(x_c, y, z; \theta_c) + L((x_c, x_m), y, z; \theta_{cm})\end{aligned} \tag{6.3}$$

where each multi-task loss $L(\cdot)$ (Eq. 6.1) is a function of different combinations of the input modalities. For example, the first term $L((x_d, x_c, x_m), \cdot)$ is a function of the dermoscopy image, clinical image, and meta-data. While the last term $L((x_c, x_m), \cdot)$ is a function of the clinical image and the meta-data (but not the dermoscopy image). We represent all parameters in the model as θ and use the subscripts to indicate the input type that updates the parameters (e.g., θ_{dm} represents parameters updated based on x_d and x_m).

At inference time, as each multi-task loss function only depends on a subset of the input types, given a specific combination of the input modalities, we can use the predictions from the classification layer that matches the available input (e.g., if only a dermoscopy image is available, use the classification layer trained only on dermoscopy images). The architecture is further defined in Fig. 6.1 and in Section 6.2.3.

During training, given a dataset of N cases, our goal is to learn the parameters θ^* of the CNN that minimizes,

$$\theta^* = \text{argmin}_\theta \sum_{i=1}^{N} \mathcal{L}(x_c^{(i)}, x_d^{(i)}, x_m^{(i)}, z^{(i)}, y^{(i)}; \theta) + \gamma||\theta|| \tag{6.4}$$

where $\mathcal{L}(\cdot)$ is defined as in Eq. 6.3, and $||\theta||$ is the L2 norm regularization term weighted by $\gamma = 0.0005$ (experimenting with other $\gamma = [0.00005, 0.0001, 0.001]$ values yielded less than 1% differences in averaged accuracy and AUROC scores). In practice, Eq. 6.4 is often minimized using gradient descent with randomly sampled mini-batches of size N_b. However, in imbalanced datasets where the frequency of the labels greatly differs, training a model on randomly sampled mini-batches with imbalanced labels can lead to a model that is biased towards the majority class, as infrequent labels contribute little to the parameter updates.

6.2.2 Mini-Batches Sampled and Weighed by Label

To address the label imbalance problem, for each mini-batch with N_b cases, we ensure there exists at least k cases that belong to each unique label. To enforce this, each mini-batch is formed by randomly sampling with replacement k cases for each unique label. This causes the model weights to be updated based on all the unique labels in each gradient descent step. As we have 24 unique labels across all categories (Fig. 6.1 L-block), this constrains our mini-batches to be of size $N_b = 24k$.

While sampling by labels improves class balance, labels within a mini-batch are still imbalanced since the category labels are not mutually exclusive, and including a case within one category, will also include its labels in all other categories. In order to further address class imbalance, we assign a higher weight to cases with labels that occur infrequently within a given mini-batch,

$$w(c)_j = \frac{\max(\mathbf{1}c)}{(\mathbf{1}c)_j} \tag{6.5}$$

where $c \in \mathbb{Z}^{N_b \times N_c}$ is a matrix representing N_b cases of 1-hot-encoded labels with N_c possible labels, $\mathbf{1} \in \mathbb{Z}^{1 \times N_b}$ is composed of all ones, $\max(\mathbf{1}c)$ returns a scalar indicating the number of cases of the most frequent label, and $(\mathbf{1}c)_j$ returns a scalar indicating the number of cases with the j-th label. Since each mini-batch has at least $k > 0$ labels, we avoid divide-by-zero errors and note how each computed weight is bound by $\left[1, \frac{N_b-(N_c-1)k}{k}\right]$. To derive the upper bound we note that the maximum value the numerator of Eq. 6.5 can take is $N_b - (N_c - 1)k$, where $(N_c - 1)k$ is subtracted since there must be at least $(N_c - 1)k$ cases with different labels in a single mini-batch (enforced in our sampling). The minimum value in the denominator of Eq. 6.5 is k (also enforced in the sampling). The lower bound is 1 since the value of the denominator in Eq. 6.5 cannot exceed the numerator.

6.2.3 Architecture to Classify, Localize, and Retrieve Images

In this section we describe the details of the layers used to form our model. We build upon a model pretrained over ImageNet [140], and remove the final output task-specific layer. We define this as our *base model*, which acts as a dense feature extractor, and outputs responses of size $h \times w \times f$ (height, width, and number of feature maps, respectively).

Classify and localize from a single modality

The following layers allow us to localize the discriminative regions in an input image, and to classify categories from a single image. For each category with l labels, we add a convolutional layer with filters of size $f \times 1 \times 1 \times l$, to the $h \times w \times f$ output of the base model. As in the work of Lin et al. [109], this layer is followed by a global spatial averaging pooling layer, where the categorical cross entropy loss (Eq. 6.1) is applied to the classification layers (Fig. 6.1 L_d, L_c). These pooled output responses classify using only a single image modality. In order to highlight the important image regions that contribute to the l-th label, we visualize (Fig. 6.4) the $h \times w$ responses (before spatial global pooling) at the l-th label (Fig. 6.1 top blue and bottom yellow blocks). Separate layers are created for the clinical and dermoscopy images with parameters θ_c, and θ_d.

Classify using image and meta-data

As the meta-data (gender, lesion location, and lesion elevation) is categorical, we one-hot encode the meta-data to produce a meta-data vector. In order to classify based on image

and meta-data, we apply a global spatial averaging pooling layer to the $h \times w$ output of the base model, apply batch normalization [78], and then concatenate the 1×1 normalized visual responses with the one-hot encoded meta-data vector. We add a convolutional layer of size $f \times 1 \times 1 \times l$ for each category, to form a classification layer (Fig. 6.1 L_{dm}, L_{cm}) used with the multi-task loss (Eq. 6.1). This is repeated for both the clinical and dermoscopy modalities to update θ_{cm}, and θ_{dm}.

Multi-modal feature vectors to retrieve and classify

We combine information from both clinical and dermoscopy images by adding a convolutional layer of size $f \times 1 \times 1 \times r$ that takes as input the global average pooled responses of the clinical and dermoscopy specific models. This produces an r-dimensional feature vector (Fig. 6.1 green bar) that is a function of both the clinical and dermoscopy image, which we use for multi-modal image retrieval (see Results). We concatenate the one-hot encoded meta-data to these visual features, add a convolutional layer for each category, and apply the multi-task loss (Eq. 6.1) to form a final classification layer (Fig. 6.1 L_{dcm}) where parameters θ_{dcm} are updated based on the dermoscopy images, clinical images, and the meta-data.

6.2.4 Inferring a Melanoma Diagnosis

As our model both directly classifies the disease diagnosis, and classifies each of the 7-point checklist criteria, there are two ways to infer a melanoma diagnosis. The first is to *directly classify melanoma* from the diagnosis category, and the second is to *infer melanoma based on the 7-point criteria* [9]. To infer melanoma based on the 7-point criteria, given predictions $\hat{z}_j^{(i)}$ for the j-th 7-point criteria of the i-th case, we compute a melanoma score $\hat{\tau}^{(i)}$, which, if exceeds a threshold t, indicates a prediction of melanoma,

$$\hat{y}_{7\text{pt}}^{(i)} = \begin{cases} \text{melanoma}, & \text{if } \hat{\tau}^{(i)} \geq t \\ \text{not melanoma}, & \text{otherwise} \end{cases} \tag{6.6}$$

$$\text{where, } \hat{\tau}^{(i)} = \sum_{j=1}^{7} \text{score}(\hat{z}_j^{(i)})$$

using a $\text{score}(\hat{z}_j)$ function that looks up the *7pt-score* from Table 6.2 that corresponds to the predicted 7-point label $\hat{z}_j^{(i)}$. The original threshold was $t = 3$ [9], which was later revised to $t = 1$ [7] in order to improve sensitivity [68]. We report results for both directly classifying melanoma, as well as inferring melanoma based on the 7-point checklist under varying thresholds in the Results section.

Table 6.1: Details of the diagnosis labels within the dataset. The *abbrev* column indicates the abbreviation for the label; *name* represents the full name of the label; and, *# cases* indicates how many unique lesions (cases) exist with the particular label. Labels with the same abbreviation are grouped in our experiments.

DIAGNOSIS (DIAG)		
abbrev.	name	# cases
BCC	basal cell carcinoma	42
NEV	blue nevus	28
NEV	clark nevus	399
NEV	combined nevus	13
NEV	congenital nevus	17
NEV	dermal nevus	33
NEV	recurrent nevus	6
NEV	reed or spitz nevus	79
MEL	melanoma	1
MEL	melanoma (in situ)	64
MEL	melanoma (less than 0.76 mm)	102
MEL	melanoma (0.76 to 1.5 mm)	53
MEL	melanoma (more than 1.5 mm)	28
MEL	melanoma metastasis	4
MISC	dermatofibroma	20
MISC	lentigo	24
MISC	melanosis	16
MISC	miscellaneous	8
MISC	vascular lesion	29
SK	seborrheic keratosis	45

Table 6.2: Details of the seven-point criteria categories within the dataset. Section headers indicate the categories. The *7pt-score* column indicates the contribution to the seven-point melanoma score by the label. See Table 6.1 for the definition of the other columns. Within a category, labels assigned the same abbreviation are grouped in our experiments.

SEVEN POINT CRITERIA			
abbrev.	name	7pt-score	# cases
1. Pigment Network (PN)			
ABS	absent	0	400
TYP	typical	0	381
ATP	atypical	2	230
2. Blue Whitish Veil (BWV)			
ABS	absent	0	816
PRS	present	2	195
3. Vascular Structures (VS)			
ABS	absent	0	823
REG	arborizing	0	31
REG	comma	0	23
REG	hairpin	0	15
REG	within regression	0	46
REG	wreath	0	2
IR	dotted	2	53
IR	linear irregular	2	18
4. Pigmentation (PIG)			
ABS	absent	0	588
REG	diffuse regular	0	115
REG	localized regular	0	3
IR	diffuse irregular	1	265
IR	localized irregular	1	40
5. Streaks (STR)			
ABS	absent	0	653
REG	regular	0	107
IR	irregular	1	251
6. Dots and Globules (DaG)			
ABS	absent	0	229
REG	regular	0	334
IR	irregular	1	448
7. Regression Structures (RS)			
ABS	absent	0	758
PRS	blue areas	1	116
PRS	white areas	1	38
PRS	combinations	1	99

6.3 Results

Our full dataset as described in Section 6.2 contains 1,011 cases. We use 413 cases to train the model (Eq. 6.4), 203 cases to validate design decisions (i.e., set hyper-parameters), and 395 cases to test and report results. Subsets were chosen to ensure a similar distribution of categories. In four cases the clinical image was missing and we used the corresponding dermoscopy image instead. All images were resized to 512×512×3.

The original dataset contains labels at the most granular level (Table 6.1 and Table 6.2). As some labels occur infrequently (e.g., two wreath vascular structure cases) and many labels have a similar clinical interpretation (e.g., types of benign nevi), we group infrequent labels with similar clinical interpretations into a single label. For example, in the diagnosis category, the NEV label groups all the nevi labels (e.g., blue nevus, clark nevus, etc) into a single label. We follow the same approach for the 7-point criteria where infrequent labels with similar clinical meaning and melanoma score contributions (i.e., a value in the *7pt-score* column in Table 6.2) are grouped. For example, within the category vascular structures, we group linear irregular and dotted labels into a single irregular label IR as the presence of either is indicative of melanoma. The final label grouping is shown in the *abbrev* column in Table 6.1 and Table 6.2.

To quantify the prediction performance of our method, for each category, we compute the prediction accuracy to indicate each category's overall performance (Table 6.3). Accuracy, however, summarizes the performance over all labels, and may hide the performance of infrequent labels. Thus we also report detailed metrics for each label (Table 6.4, 6.5, 6.6).

We first report results using the most frequent training set labels as the test predictions in order to compute baseline results in the context of an imbalanced dataset. Table 6.3 (experiment *frequent*) shows that this simple approach yields an average accuracy of 61.8%, and thus model performance should be considered relative to this baseline.

6.3.1 Model Training

Our experiments use Inception V3 [153], pretrained over ImageNet [140] as our *base model*. We replace the class-specific layer with a trainable layer for each loss function as described in Section 6.2.3 and illustrated in Fig. 6.1. We augment the training images in real-time with flips, rotations, zooms, and height and width shifts. To train, we freeze all pretrained parameters, and train with a learning rate of 0.001 for 50 epochs, then reduce the learning rate to 0.0001 for 25 epochs, unfreeze the deepest frozen "inception block", and repeat for 25 epochs until all layers are unfrozen up to the second "inception block". Finally, we train for 25 epochs on non-augmented data, for a total of 300 epochs. We use Keras [37] with TensorFlow [1] to create and optimize our models using stochastic gradient descent with a decay of 1e-6 and momentum of 0.9. We observed that even though our model was trained with multiple loss functions (Eq. 6.3), it consistently reduced the loss over the training data.

Table 6.3: The accuracy of each of the seven-point criteria and diagnosis. The column *avg.* averages the accuracy over each row.

Experiment	BWV	DaG	PIG	PN	RS	STR	VS	DIAG	avg.
frequent	81.0	44.8	56.5	39.5	73.2	65.1	79.2	55.4	61.8
x-unbalanced	87.6	56.7	65.6	68.1	78.2	75.9	81.3	68.4	72.7
x-balanced	87.3	60.3	64.8	68.9	78.2	75.7	81.5	70.9	73.4
x_c	79.2	52.7	56.5	57.0	71.6	60.3	75.2	60.0	64.1
x_c+x_m	77.7	51.9	59.2	59.5	72.9	62.8	76.7	61.5	65.3
x_d	85.8	60.8	62.8	69.4	77.5	71.4	80.3	71.9	72.5
x_d+x_m	85.1	59.7	63.3	69.4	76.7	74.2	81.5	73.4	72.9
x-combine	87.1	60.0	66.1	70.9	77.2	74.2	79.7	74.2	**73.7**
x_d+x_c-retrieve	86.8	56.7	62.8	65.3	78.0	73.4	81.0	71.1	71.9
Ngiam [130]	83.0	59.2	61.3	65.6	73.9	69.4	75.7	70.6	69.8
x_c	77.5	50.6	52.9	56.5	67.8	59.7	75.9	58.2	62.4
x_d	82.5	60.5	63.3	67.8	69.6	71.1	72.7	66.8	69.3

6.3.2 Unbalanced vs. Balanced Training

We compare the performance of a model trained on balanced data by first training a model using random mini-batches with uniform class weights, and report results under the experiment name *x-unbalanced*. Unless otherwise stated, results are computed using the predictions that are a function of the entire input (i.e., Fig. 6.1 L_{dcm}). We compare *x-unbalanced*, to the same model trained on mini-batches sampled and weighted by label (described in Sec 6.2.2) using the experiment name *x-balanced*. When training with balanced mini-batches, we observe a small increase in overall accuracy; however, as noted earlier, accuracy does not well highlight improvements made to classifying infrequent labels. The averaged metrics across all labels increase for the 7-point checklist (Table 6.4) and diagnosis (Table 6.5), when compared to the model trained without balancing the classes (*x-unbalanced*). Notably, *x-balanced* improves the sensitivity and precision of detecting irregular vascular structure (VS IR in Table 6.4) from 0% in *un-balanced* for both, to 10% and 60%, respectively. A similar performance increase is seen in the sensitivity (5.3% to 21.1%) and precision (33.3% to 50%) of detecting seborrheic keratosis (SK in Table 6.5). To compare the performance of the imbalanced experiment (i.e., *x-unbalanced*) with the balanced experiment (i.e., *x-balanced*), we apply a Friedman test [60] using AUROC scores for each category, where the AUROC scores are averaged within each category, and obtained a statistically significant difference between the two models ($p = 0.0047$).

Table 6.4: The seven-point criteria results. Columns indicate the seven-point criteria, separated by the labels that belong within each criteria. The final *avg.* column is the result averaged over the entire row. Each row represents an experiment, divided into results for sensitivity (*sens.*), specificity (*spec.*), precision (*prec.*), and area under the receiver operating characteristic curve (*auroc.*). Label abbreviations are defined in Table 6.2.

7pt criteria		BWV		DaG			PIG			PN			RS		STR			VS			
Experiment.	met.	ABS	PRS	ABS	REG	IR	ABS	REG	IR	ABS	TYP	ATP	ABS	PRS	ABS	REG	IR	ABS	REG	IR	Avg.
x-unbalanced	sens.	96.6	49.3	34.0	59.3	67.8	83.0	6.2	57.3	78.8	77.4	35.5	95.5	31.1	98.1	36.4	34.0	98.7	23.1	0.0	55.9
	spec.	49.3	96.6	92.2	72.2	67.4	53.5	99.4	80.1	80.8	75.5	93.7	31.1	95.5	47.8	98.6	94.0	22.0	97.1	100.0	76.1
	prec.	89.0	77.1	59.6	47.6	62.8	69.8	60.0	56.8	72.8	64.9	63.5	79.1	71.7	77.8	76.2	64.0	82.8	54.5	0.0	64.7
	auroc	87.0	87.0	72.3	72.6	76.4	77.4	67.2	78.1	87.8	83.6	78.6	79.9	79.9	84.2	87.8	78.3	82.1	81.8	73.4	79.8
x-balanced	sens.	92.5	65.3	43.0	66.1	66.1	73.5	16.7	67.7	78.2	76.0	41.9	84.1	62.3	90.7	43.2	50.0	96.8	30.8	10.0	60.8
	spec.	65.3	92.5	89.8	75.1	73.4	64.5	98.3	73.4	81.6	77.9	92.1	62.3	84.1	63.8	97.4	87.7	31.7	95.6	99.5	79.3
	prec.	91.9	67.1	58.9	53.1	66.9	72.9	57.1	53.8	73.5	66.9	61.9	85.9	58.9	82.3	67.9	56.0	84.4	51.6	60.0	**66.9**
	auroc	87.5	87.5	73.0	76.5	78.0	78.8	75.2	79.4	88.6	83.6	78.9	83.5	83.5	84.9	87.1	78.7	85.0	84.0	76.1	81.6
x-combine	sens.	89.4	77.3	47.0	67.8	62.1	77.6	29.2	59.7	77.6	78.1	48.4	81.3	66.0	86.0	54.5	51.1	92.3	42.3	13.3	**63.2**
	spec.	77.3	89.4	87.8	72.6	78.9	65.1	94.2	80.1	85.8	78.7	90.7	66.0	81.3	67.4	96.0	85.7	45.1	92.4	97.5	**80.6**
	prec.	94.4	63.0	56.6	51.3	70.5	74.2	41.2	57.8	78.1	68.3	61.6	86.7	56.5	83.1	63.2	52.7	86.5	45.8	30.8	64.3
	auroc	89.2	89.2	74.1	76.5	79.9	79.0	74.9	79.0	89.9	84.2	79.9	82.9	82.9	86.1	87.0	78.9	86.2	85.5	76.1	**82.2**
x_d+x_c retrieve	sens.	91.9	65.3	36.0	63.6	63.8	77.1	18.8	54.0	73.1	71.9	41.9	88.9	48.1	85.6	52.3	50.0	92.7	40.4	30.0	60.3
	spec.	65.3	91.9	91.5	70.0	71.1	59.3	96.0	76.8	79.9	77.5	89.1	48.1	88.9	66.7	95.2	86.0	45.1	93.9	97.5	78.4
	prec.	91.9	65.3	59.0	47.5	64.2	71.1	39.1	51.5	70.4	65.2	54.2	82.4	61.4	82.7	57.5	52.8	86.6	50.0	50.0	63.3

Table 6.5: The results for the diagnosis category, and for melanoma prediction based on the predicted seven-point scores. The final columns *Mel7* shows the results using the scores from the predicted seven-point checklist to predict melanoma using two common thresholds, $t = 1$ and $t = 3$.

		DIAG						Mel7	
Experiment	met.	BCC	NEV	MEL	MISC	SK	Avg.	t=1	t=3
x-unbalanced	sens.	25.0	94.1	44.6	35.0	5.3	40.8	90.1	47.5
	spec.	98.4	50.6	92.2	98.0	99.5	87.7	40.1	87.4
	prec.	40.0	70.3	66.2	66.7	33.3	55.3	34.1	56.5
	auroc	92.2	87.7	83.2	86.3	88.4	87.6	76.8	
x-balanced	sens.	25.0	91.3	55.4	42.5	15.8	46.0	96.0	69.3
	spec.	98.9	62.5	88.4	97.2	99.7	89.4	33.0	78.9
	prec.	50.0	75.2	62.2	63.0	75.0	65.1	33.0	53.0
	auroc	89.2	88.1	84.2	86.8	90.4	87.7	81.7	
x-combine	sens.	62.5	88.6	61.4	47.5	42.1	**60.4**	96.0	69.3
	spec.	97.9	71.6	88.8	97.5	99.5	**91.0**	36.1	77.6
	prec.	55.6	79.5	65.3	67.9	80.0	**69.6**	34.0	51.5
	auroc	92.9	89.7	86.3	88.3	91.0	**89.6**	81.6	
x_d+x_c retrieve	sens.	37.5	87.2	59.4	42.5	36.8	52.7	94.1	73.3
	spec.	97.9	69.9	87.8	97.2	98.1	90.2	36.1	78.6
	prec.	42.9	78.3	62.5	63.0	50.0	59.3	33.6	54.0

6.3.3 Performance Based on Input

To examine the classification performance as a function of the input modalities, we report the average accuracy using the predicted responses from different classification layers (L blocks in Fig. 6.1). We see the average classification accuracy using the clinical images and meta-data (experiment x_c and x_c+x_m in Table 6.3) is much lower than when using the dermoscopy images and meta-data (experiment x_d and x_d+x_m). Dermoscopy images likely yield higher classification accuracy since the 7-point checklist was designed for criteria visible under dermoscopy. The classification layer that uses clinical, dermoscopy, and meta-data together yields the highest average accuracy. However, we note including clinical images gives relatively small improvements over using dermoscopy images alone, and that this improvement may be partly due to the additional layer that joins the two modalities. Our observations differ from those reported by Ge et al. [63], which showed larger accuracy improvements when incorporating clinical and dermoscopy images into a single model, as well as similar diagnosis performance for each modality. We note Ge et al. [63] report results over a larger dataset, which may in part explain our different observations. Our results, separated by input modalities, illustrate that our approach degrades gracefully with missing data, making it applicable to scenarios where only partial patient data is available.

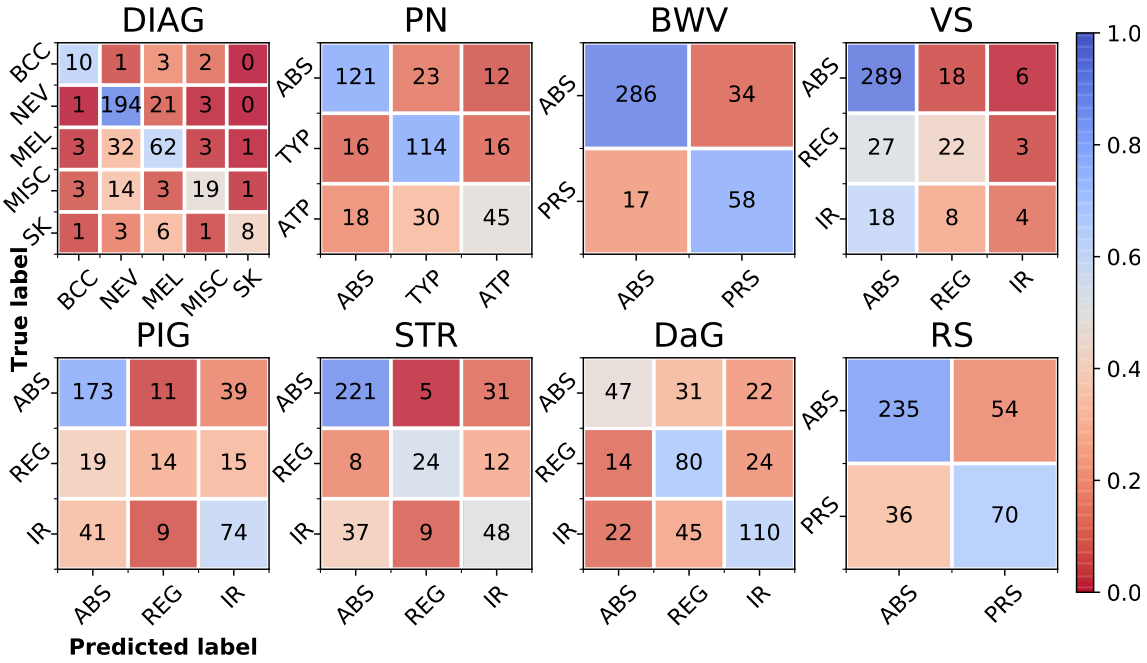


Figure 6.2: Confusion matrices for each category using the test set predictions from our proposed model. The y-axis indicates the ground truth labels. The x-axis indicates the model's predicted labels. Numbers in each entry represent the number of cases classified as such. Colours indicate the percentage of each label in each entry, normalized by the total number of true labels.

6.3.4 Other Multi-Modal Approaches

Our approach of using multiple losses that are a function of different input modalities (Eq. 6.3), differs from other multi-modal approaches such as the work by Ngiam et al. [130], which randomly sets some input modalities to zero during training. We perform an additional experiment based on Ngiam et al.'s work [130], where on average we set a single input modality to zero in 75% of the samples within a mini-batch. The other 25% includes all three modalities. We remove all loss functions except for L_{dcm} (Fig. 6.1), and repeat the *x-balanced* experiment. We report test results in Table 6.3 for predictions based on all three modalities, only clinical images, and only dermoscopy images. We obtain consistently higher averaged accuracy in our proposed approach for each type of input. One possible reason for our improvement is that Ngiam et al. [130] learn a model that is robust to missing data, which may compete with learning disease patterns specific to a single modality. Whereas our approach may learn patterns specific to each modality, as the loss functions are trained on each individual modality.

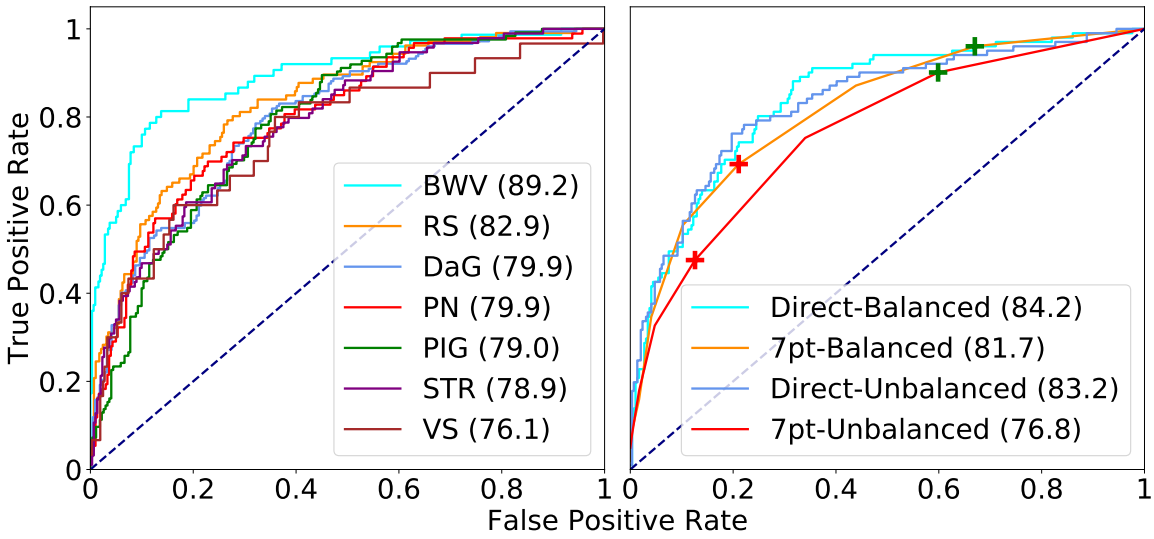


Figure 6.3: *(Left)* One-vs-all ROC curves for each label in the 7-point criteria that contribute to melanoma. *(Right)* Melanoma ROC curves comparing direct melanoma classification with inference via the 7-point checklist, using unbalanced and balanced training procedures. The green and red cross indicates the threshold of 1 and 3, respectively, used in Eq. 6.6.

6.3.5 Combining Classification Layers' Predictions

We also report results from averaging the predicted probabilities of the three classification layers that are a function of dermoscopy images (Fig. 6.1 L_d, L_{dm}, L_{dcm}) into a final prediction (experiment name *x-combine*). While this results in a minor decrease to the average precision over the 7-point checklist when compared to *x-balanced*, average sensitivity increases, and all metrics are increased in the diagnosis category (Table 6.5). This approach of combining multiple classification layers is analogous to averaging the predictions from multiple independent neural networks; however, our model shares most layers. We use the *x-combine* predictions to form the confusion matrices in Fig. 6.2 and the ROC curves in Fig. 6.3 (*left*).

6.3.6 Inferring Melanoma

As noted in the Methods section, we can infer a melanoma diagnosis by either direct classification, or via the 7-point checklist (Eq. 6.6). Fig. 6.3 (*right*) shows the ROC curve from directly diagnosing melanoma, and from thresholds based on the 7-point score (i.e., t in Eq. 6.6). We see that directly classifying melanoma yields a higher AUROC score than the predicted 7-point scores. However, at high sensitivity levels, the performance of both approaches are similar. In addition, the direct classification AUROC score comes at the cost of a less interpretable model, as this ROC curve is based on thresholding probabilities for a binary decision (melanoma vs all), which is clinically less interpretable than the 7-point scores. We highlight that our approach outputs both results, and either diagnoses approach can be

Table 6.6: Related works separated by category and labels. We report the aggregated metrics used in the original works. The *rep* column indicates if we could replicate the same training/test images and report a direct comparison.

	rep.	category (labels)	acc.	sens.	spec.	prec.	auroc
Sadeghi [141]	✗	STR	76.1	76.0*	-	74.2*	85.0*
ours		(ABS, REG, IR)	74.2	74.2*	74.9*	73.6*	84.5*
Wadhawan [165]	✗	BWV	-	79.5	79.2	-	-
ours		(ABS, PRS)	87.1	77.3	89.4	63.0	89.2
Wadhawan [165]	✗	RS	-	64.2	67.9	-	-
ours		(ABS, PRS)	77.2	66.0	81.3	71.6	82.9
Menegola [122]	✓	DIAG (BCC,	-	-	-	-	84.5
ours		MEL, Other)	80.8	64.9	84.8	74.6	88.5

*Metric averaged by weighted sample. Other metrics are unweighted averages, except for the binary cases of *sens*, *spec*, and *prec* .

used. Finally, Fig. 6.3 (*right*) also shows that our *x-balanced* training improves melanoma detection for both approaches.

6.3.7 Works Using the Same Data

Comparing with other approaches is challenging as often different subsets of the data are used from various sources, with multi-class labels grouped to form binary problems. We compared with works that used the same dataset, and that reported the same class labels as our work. This is reported in Table 6.6, along with a checkmark indicating if the exact subsets of the data used in this work was publicly available, allowing for a direct comparison. Sadeghi et al. [141] classified absent, regular, and irregular streaks using 945 images, of which 745 are from the same dataset as our work. Wadhawan et al. [165] used 347 "low difficulty" images from the same dataset as our work, and we compare with the two categories that we both report binary labels on. Our results do not exclude challenging images and do not rely on lesion segmentations. Menegola et al. [122] make the image names and cross-validation folds publicly available. We run new experiments using the same image names, perform 5 rounds of 2-fold cross validation based on their provided folds, and modify our diagnosis loss function to follow their 3-class experiment (melanoma vs. basal cell carcinoma vs other benign lesions). We follow the same training and inference procedure as *x-combine* and compare with their top performing approach. Table 6.6 suggests our model achieves results comparable to state-of-the-art among various categories.

6.3.8 Localization

With the goal of providing a model whose classification may be interpretable by humans, we visualize the areas of the image that contribute to the predicted label by viewing the learned $h \times w$ responses that correspond to the l-th label. Given an image, we re-size the

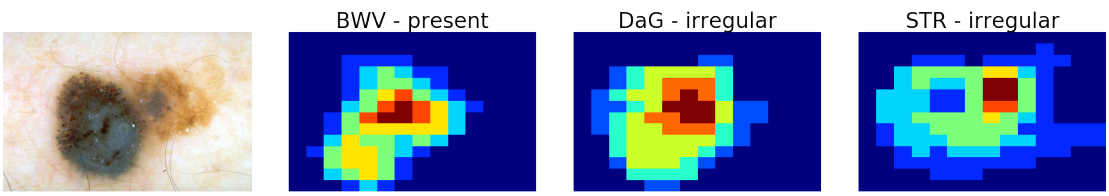


Figure 6.4: Learned responses localize the image areas that contribute to the specific class label for a given input image.

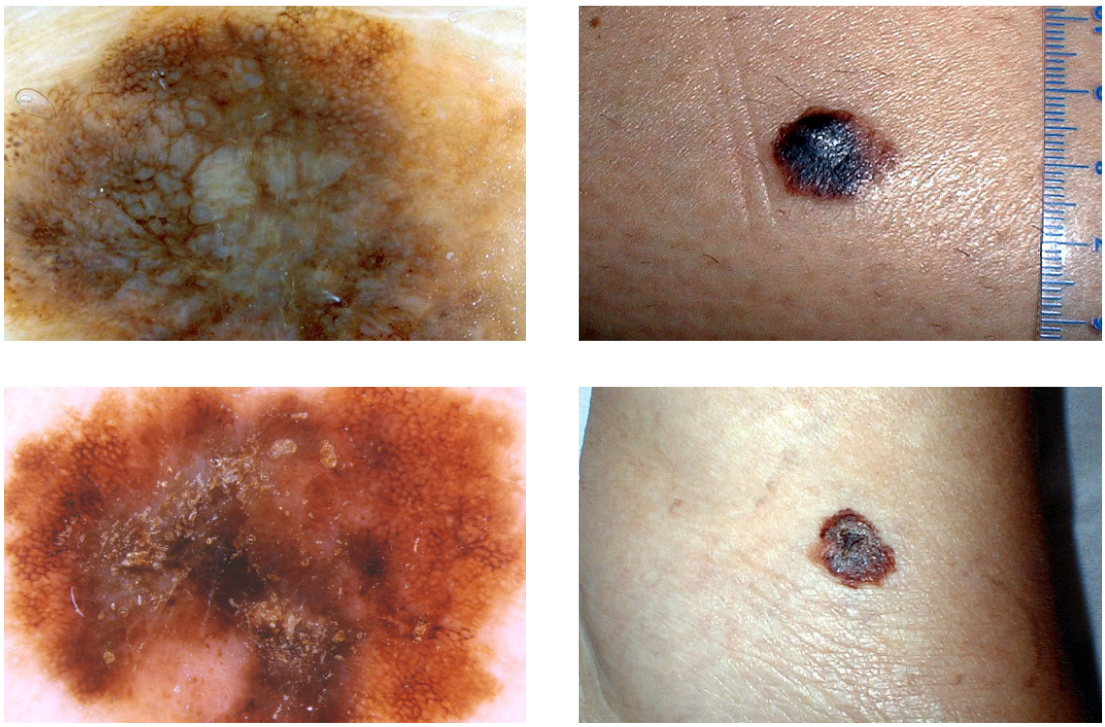
$h \times w$ responses (e.g., x_d responses are represented by the top blue box in Fig. 6.1) to match the size of the original image, and show the response for select labels in Fig. 6.4. By visualizing those areas that influence the classification, users can check for the presence of these criteria in the detected areas and adjust their confidence in the machine's prediction accordingly. Ge et al. [63] used a similar approach based on class activation maps [178], to visualize the diagnosis category, while here we show localized results for clinical criteria.

6.3.9 Image Retrieval

We demonstrate our approach is able to retrieve clinically similar images with respect to the 7-point criteria and diagnosis (Fig. 6.5). For each image, we extract the r-dimensional responses (Fig. 6.1 green rectangle) that are a function of both the clinical and dermoscopy images (Sec. 6.2.3). For each test case, we find the training case feature vector with the lowest cosine distance, and use the known training labels as our predictions (experiment x_c+x_d-*retrieve*). In Chapter 4, Kawahara et al. [92] used a similar approach to retrieve a path of visually similar images; whereas, this works learns a new compact multi-modal feature vector. This image retrieval approach achieves comparable averaged results (Tables 6.3, 6.4, 6.5) with the classification based approach. However, image retrieval has the additional advantage of allowing users to infer labels from expertly labeled images, rather than relying on a black box classification system, and may prove more interpretable than classification or localization approaches. We note how our multi-modal r-dimensional feature vector (Fig. 6.1 green bar) retrieves multiple modalities with a *single* vector, and that our loss function (Eq. 6.3) learns compact feature vectors that represents several clinically relevant criteria.

6.4 Conclusions

We propose a neural network designed for multi-modal images and meta-data, that classifies all seven-point checklist criteria and skin lesion diagnosis within a single optimization (multi-task). Our architecture uses multiple loss functions to handle combinations of the input modalities, and at inference time is capable of making predictions with missing data. Further, our architecture is capable of localizing discriminative information, and produces



	DIAG	PN	BWV	VS	PIG	STR	DaG	PS
top	MEL	ATP	PRS	ABS	ABS	IR	IR	ABS
bottom	MEL	ATP	PRS	IR	IR	IR	IR	ABS

Figure 6.5: Dermoscopy (*left*) and clinical (*right*) images of a lesion from the *test* set (*top row*) and the most visually similar image in the training set (*bottom row*). The table labels correspond to the top and bottom row, respectively.

feature vectors useful for image retrieval of clinically similar images. We observe that, for some criteria, our model is unable to distinguish among the labels (e.g., model almost always predicts absent vascular structures). We see these as active areas for improvement and hope for further progress by the research community with the release of this dataset.

Chapter 7

Fully Convolutional Neural Networks to Detect Dermoscopic Criteria

7.1 Introduction

In order to distinguish melanoma from benign lesions, dermatologists often rely on using melanoma-specific image cues to aid in their diagnoses. Dermoscopy images, which are captured with a dermatoscope, offer a magnified view of the skin lesion and allow dermatologists to visualize structures within the lesion that may indicate melanoma [99]. For example, as discussed in Chapter 6, the 7-point checklist [9] is a scoring system that checks for the presence of visual cues (e.g., streaks) in dermoscopy images, and assigns a numerical score that, if exceeded, may indicate melanoma. This helps give dermatologists an objective criteria on which to base their diagnosis.

7.1.1 Detecting Dermoscopic Criteria

Many groups have studied how to detect and classify dermoscopic criteria from dermoscopy. Celebi et al. [34] detected the blue-whitish veil in dermoscopy images. They formed a feature vector using colour and texture based features from patches of pixels, and used a decision tree to classify the patch. Sadeghi et al. [141] proposed geometric, structural, orientation, and chromatic features to capture the properties of streaks. Combined with colour and texture based features, they classified absent, regular, and irregular streaks. Mirzaalian et al. [125] modeled the tubular properties of streaks with a Hessian based tubular filter. They computed a feature vector by measuring the detected flux through multiple iso-distance contours to the lesion's boundary, and trained a support vector machine (SVM) classifier to classify absent, regular, or irregular streaks. Barata et al. [18] proposed using directional filters in dermoscopy images to detect the presence of pigment networks. They formed

feature vectors used for classification based on the density and distribution properties of the detected pigment networks.

7.1.2 Deep Learning to Segment and Classify Skin Lesions

Previous work has shown CNNs to be useful for both skin lesion segmentation and classification tasks [40, 42, 85, 90, 113, 176]. CNNs have stacked layers of convolution filters with, commonly, millions of free parameters (also called weights) that learn to represent the data at different levels of abstraction [102]. These free parameters are often learned through a training process where example images and their corresponding labels (e.g., diagnoses or segmentation masks) are used to update the CNN's free parameters such that the network learns to produce outputs that match the labels. In order to learn free parameters that give a useful abstraction of the data, CNNs often are trained on large datasets of images. As existing skin datasets are relatively small, a common approach [40, 42, 85, 90, 113, 176] is to use the parameters of a CNN already trained over a larger dataset [140]. This leverages the useful data abstractions learned over larger datasets for smaller datasets.

7.1.3 Sørensen-Dice-F_1 Score as a Loss Function

Training a CNN typically requires minimizing a loss function. As the model's parameters are updated to minimize the loss, the choice of the loss function influences the resulting trained model. The Sørensen-Dice coefficient or F_1 score has been proposed as a loss function for imbalanced datasets [123, 133, 149]. We note that the Sørensen-Dice coefficient and the F_1 score are equivalent (discussed in Section 7.2.5). Pastor-Pellicer et al. [133] proposed the negative F_1 score as a loss function for neural networks in order to clean and enhance ancient document images. Milletari et al. [123] proposed using the Sørensen-Dice coefficient as the loss function for a neural network designed for volumetric segmentation. Sudre et al. [149] proposed using the Sørensen-Dice coefficient weighted by the size of the object within the image as the neural network loss function for 2D and 3D segmentation.

7.1.4 Skin Lesion Datasets and Competitions

Korotkov et al. [99] noted that one of the major limitations of computerized skin lesion analysis research is the lack of standardized skin lesion datasets, and that the "creation of such a dataset is of utmost importance for future development of this field". Fortunately, since that review, new skin lesion datasets have become available such as DermoFit [12], and PH^2 [120]. More recently, the International Skin Imaging Collaboration (ISIC), in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI), began hosting a skin lesion analysis competition [41, 66]. In addition to providing a standardized dataset, this public competition offers standard evaluation procedures and metrics in order to benchmark lesion segmentation, dermoscopic criteria detection, and lesion classification

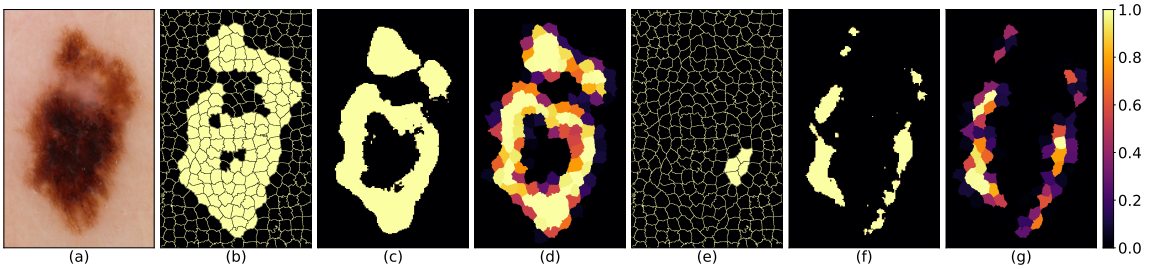


Figure 7.1: Superpixels to segmentations, and segmentations to superpixels. (a) The original image. Expertly annotated (b) pigment-network and (e) streak superpixels converted to binary segmentations, overlaid with superpixels. Pixel-wise (c) pigment-network and (f) streaks CNN predictions. CNN predictions converted to (d) pigment-network and (g) streak superpixels. Images shown here are cropped around the lesion for visualization purposes.

approaches. In this work, we focus on *Part 2: Dermoscopic Feature Classification Task* of the 2017 ISIC-ISBI challenge [41]. This task involves classifying superpixels that may contain the presence of dermoscopic criteria.

7.1.5 Contributions

In the previous chapter (Chapter 6), our goal was to classify the labels (e.g., absent, regular, irregular) for multiple dermoscopic criteria, where for each criterion a label was assigned to the entire image. In this chapter, our goal is to detect (i.e., localize and classify) the presence or absence of multiple dermoscopic criteria, where a binary label for each dermoscopic criterion is assigned to each pixel (or superpixel).

In the following sections we detail our proposed approach that reformulates the superpixel classification task as a segmentation problem, and finetunes a pretrained CNN to detect pixels that contain the studied dermoscopic criteria. Our CNN architecture is modified for semantic segmentation, and is trained to minimize a negative multi-label fuzzy Sørensen-Dice-F_1 score, where the score is computed over partitions of the mini-batch. This approach ranked first place in the *2017 ISIC-ISBI Part 2* task [41], which used the area under the receiver operator characteristic curve (AUROC) to evaluate submissions. We discuss the limitations of the metrics used to rank the challenge entries, and show two simple baseline methods that empirically outperform all entries when ranked by the current and past challenge metrics. We propose to use a fuzzy Jaccard Index that ignores the empty set (i.e., when neither predicted nor ground contain positive values) to rank model performance, rather than AUROC. We plan to publicly release our trained model along with the code used to create and train the model.

7.2 Methods

Given a dermoscopy image x, and a corresponding superpixel labelling mask s, our task is to predict the set of dermoscopic criteria labels z that belong to each superpixel. The label z_i assigns to the i-th superpixel s_i the following K potentially overlapping dermoscopic criteria: *pigment network*; *negative network*; *milia-like cysts*; and *streaks*. These are represented as binary vectors of length $K = 4$. For example, $z_i = [1, 0, 0, 1]$ indicates the presence of both *pigment network* and *streaks* in the i-th superpixel.

7.2.1 Motivations to Segment Instead of Label Superpixels

While labelling superpixels is a convenient way to gather ground truth data from human clinicians as it avoids a detailed per-pixel labelling, individual superpixel labelling is less desirable for machine classification tasks for the following two reasons. Firstly, by considering each superpixel individually, the machine classifies based only on the local context available within a superpixel, and ignores surrounding context such as location relative to the entire lesion (e.g., dermoscopic criteria commonly occur within or near the border of the lesion). Secondly, many state-of-the-art approaches for classification rely on a deep learning framework [140]. Classifying individual superpixels within a deep learning framework is challenging, as typical deep learning frameworks expect a fixed sized rectangular input, whereas, individual superpixels are of varying size and have non-rectangular shapes. Further, converting to a more conventional deep learning approach allows us to take advantage of neural networks pretrained over larger datasets.

7.2.2 Superpixels to Segmentations

As previously motivated, rather than treating this as a superpixel classification problem, we instead model this as a multi-label segmentation task. We convert the superpixels s and corresponding dermoscopic criteria labels z into a 3D volume $m \in \mathbb{Z}^{K \times W \times H}$, where K indicates the number of labels, and the width W and height H correspond to the spatial dimensions of the input image x (Fig. 7.1a). Specifically, we assign each element within m a binary label z_{ik} to indicate the presence or absence of the k-th dermoscopic criteria at a particular element m_{kwh},

$$(w, h) \in s_i \implies (m_{kwh} = z_{ik}) \tag{7.1}$$

where z_{ik} represents the k-th label for the i-th superpixel, and the superpixel s_i is composed of (w, h) spatial locations that index into the spatial locations of m. This representation captures the spatial dependencies among superpixels, and allows us to efficiently leverage pretrained CNNs.

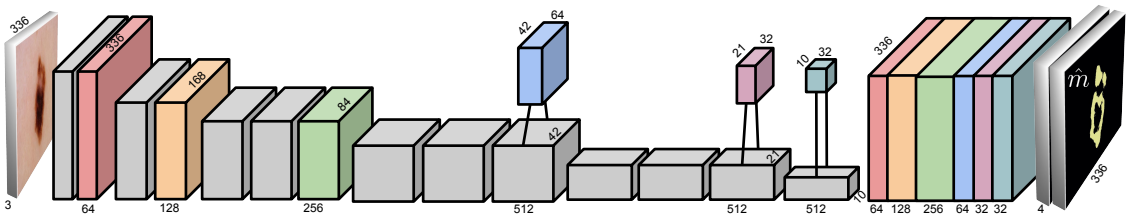


Figure 7.2: The CNN used to detect dermoscopic criteria. Feature maps from six layers are resized to match the spatial dimensions of the input and concatenated together. The colours indicate the selected layers that correspond to the concatenated block. We add additional convolutional layers to the deeper layers in order to reduce the number of feature maps (*floating blocks*). A final layer is added to represent each of the dermoscopic criteria.

7.2.3 Segmentations to Superpixels

While our CNN produces segmentations/pixel predictions (Fig. 7.1*c,f*), our final task is to assign a set of labels to each superpixel. We convert the predicted segmentation mask $\hat{m} \in \mathbb{R}^{K \times W \times H}$ back to a predicted superpixel labelling $\hat{z}$ (Fig. 7.1*d,g*) by assigning to the k-th label of the i-th superpixel the average probabilities predicted within the i-th superpixel location, i.e.,

$$\hat{z}_{ik} = \frac{1}{|s_i|} \sum_{w,h \in s_i} \hat{m}_{kwh} \tag{7.2}$$

where $|s_i|$ indicates the number of pixels in the superpixel s_i, and $\hat{m}_{kwh}$ is the predicted probability of the k-th label at the (w, h) spatial location.

7.2.4 CNN Architecture

We extend VGG16 [147], a CNN pretrained over ImageNet [140], using a similar semantic segmentation architecture as proposed by Long et al. [112]. We remove the fully-connected layers of VGG16, and resize selected responses/feature maps throughout the network (see Fig. 7.2 for selected layers) to match the sized of the input image using bilinear interpolation. These selected resized feature maps are concatenated, allowing us to directly consider feature maps from several network layers. This design is motivated by our observation that the appearance of dermoscopic criteria are subtle, and may be represented in shallower layers with higher spatial resolutions. However, concatenating these resized responses from several layers exceeds the memory available on modern GPUs. To lower the GPU memory requirements, and to give emphasis to feature maps from shallower layers, we reduce the number of concatenated feature maps from layers with 512 feature maps by adding additional convolutional layers with filters of size $512 \times 1 \times 1 \times F$, where F is either 64 or 32 depending on the layer (Fig. 7.2 provides details). This reduces GPU memory requirements, giving more emphasis to shallower layers, while still considering information found

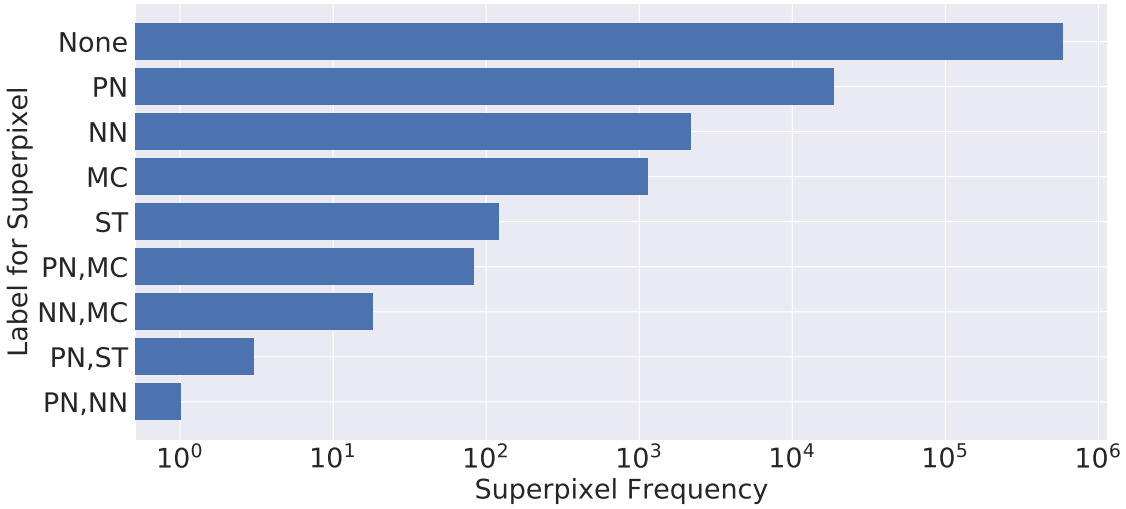


Figure 7.3: The distribution of the superpixel labels over the ISIC-ISBI 2017 test set. The *x-axis* shows the number of superpixels with a given label on a log scale, which illustrates the imbalanced data. The *y-axis* shows the labels, and is expanded to show the frequency of superpixels that are assigned multiple labels. We see that most labeled superpixels have a single label (e.g., pigment network *PN* occurs most frequently on its own), but a single superpixel can contain multiple labels (e.g., negative network *NN* and milia-like *MC* occur within the same superpixel). The majority of superpixels contain no label (*None*). Some labels do not occur within the same superpixel (e.g., streaks *ST* never occurs with *NN*) and are not shown here.

in deeper layers. Our final concatenated layer is of size $W \times H \times 576$, which matches the spatial dimensions of the input image x.

Our final layer adds an additional convolutional layer with a filter of size $576 \times 1 \times 1 \times K$ to the concatenated block. This represents our output (i.e., segmentation) for each of the K dermoscopic criteria. A sigmoid activation function is applied element-wise to scale the output between 0 and 1. These K additional *channels* represent the labels for the K types of dermoscopic criteria. Note that we do not apply the softmax activation function to this final layer, since the dermoscopic criteria can overlap.

7.2.5 Negative Multi-Label Sørensen-Dice-F_1 Loss Function

The dermoscopic criteria labels z are heavily imbalanced in favour of the background, and even among the criteria, some types occur much more frequently than others. For example, in the *ISIC-ISBI Part 2* challenge training data, there are approximately 55× more pixels labelled as *pigment network*, than *negative network* (see Fig. 7.3 for the distribution of labels). Additionally, many images contain no positive instances of a specific class. We consider data imbalance from three perspectives: *pixel-imbalance*, where the background pixels dominate the foreground pixels; *class-imbalance*, where some classes occur more frequently than

others; and, *sample-imbalance*, where many samples contain no positive instances. In order to encourage the CNN to be sensitive to dermoscopic criteria and address pixel-imbalance, we base our loss on the Sørensen-Dice-F_1 score. The F_1 score for two multi-dimensional arrays $\hat{a}, a$ with n elements, where $\hat{a}_i, a_i \in [0, 1]$, is defined as,

$$D(\hat{a}, a) = \frac{2 \cdot TP(\hat{a}, a) + \alpha}{2 \cdot TP(\hat{a}, a) + FP(\hat{a}, a) + FN(\hat{a}, a) + \beta} \tag{7.3}$$

where fuzzy true positives $TP(\hat{a}, a) = \sum_i^n (\hat{a}_i \cdot a_i)$, false positives $FP(\hat{a}, a) = \sum_i^n (\hat{a}_i \cdot (1 - a_i))$, and false negatives $FN(\hat{a}, a) = \sum_i^n ((1 - \hat{a}_i) \cdot a_i)$ are computed [133]. Setting $\beta > 0$ prevents divide-by-zero errors and α controls the score returned when neither the ground truth nor the predicted labels have any positive values. Equation 7.3 can be simplified and rewritten into an equivalent form more recognizable as the Sørensen-Dice coefficient,

$$D(\hat{a}, a) = \frac{2 \cdot \sum_i^n (\hat{a}_i \cdot a_i) + \alpha}{\sum_i^n (\hat{a}_i + a_i) + \beta} \ . \tag{7.4}$$

The loss function to train a CNN is computed over mini-batches $\hat{M} \in \mathbb{R}^{B \times K \times W \times H}$, where B is the number of mini-batch samples (e.g., $\hat{m} \in \hat{M}$ is a single sample). Given the predicted $\hat{M}$ and true M mini-batch segmentations, we train the CNN to minimize a negative multi-label Sørensen-Dice-F_1 score,

$$\mathcal{L}(\hat{M}, M) = 1 - D^*(\hat{M}, M) \tag{7.5}$$

where $D^*(\hat{M}, M)$ computes the Sørensen-Dice-F_1 score over a mini-batch. $D^*(\hat{M}, M)$ can take different forms by computing $D(\cdot, \cdot)$ over various mini-batch partitions. For example, if $D^*(\hat{M}, M) = D(\hat{M}, M)$, we compute a *single* Sørensen-Dice-F_1 score for the entire mini-batch, which addresses pixel-imbalance. However, class-imbalance can cause the model to be biased towards the prevalent class label, which can result in the model ignoring infrequent class labels. To balance infrequent class labels, an intuitive choice which avoids explicit class re-weighting (as in [149]) is to compute the Sørensen-Dice-F_1 score over each of the K channels, and over each of the B mini-batch samples,

$$D^{B,K}(\hat{M}, M) = \frac{1}{B \cdot K} \sum_{b=1}^{B} \sum_{k=1}^{K} D(\hat{M}^{b,k,:,:}, M^{b,k,:,:}) \tag{7.6}$$

where $M^{b,k,:,:}$ represents a 2D array that corresponds to the b-th sample of the k-th channel. Setting $\alpha, \beta = 1$ avoids divide by zero errors, and returns a score of 1 when both the predicted and ground truth labels are all zeros (loss = 0 Eq. 7.5). However, in datasets where a large proportion of samples contain no positive labels (i.e., sample-imbalance), this can bias the classifier to learn to only predict background labels. Setting $\alpha = 0$ and $\beta = 1$ returns a score of 0 (loss = 1) when both the predicted and ground truth are all zero. While

this no longer encourages the model to learn to predict all background values, it considers all negative samples as an error regardless of the predictions, which prevents the model from learning using the negative samples. In order for the model to learn from negative samples, and to account for sample and class-imbalance without explicit re-weighting, for each channel, we compute a Sørensen-Dice-F_1 score over the entire B samples within the mini-batch,

$$D^K(\hat{M}, M) = \frac{1}{K} \sum_{k=1}^{K} D(\hat{M}^{:,k,:,:}, M^{:,k,:,:}) \tag{7.7}$$

where $M^{:,k,:,:}$ represents a 3D array composed of the k-th channel of all B samples within a mini-batch. Cases when the entire ground truth channel is composed of all negative samples will occur less frequently since B samples are considered simultaneously. Thus, computing the Sørensen-Dice-F_1 score for each mini-batch channel (rather than for each sample) allows negative samples to contribute to the learning without dominating the loss function.

7.2.6 Training and Augmented Data with Over-Sampled Classes

We train our CNN by minimizing Eq. 7.5 using the Adam optimizer [95] with a learning rate of 0.00005. Our models were built and optimized using Keras [37] with TensorFlow [1]. While VGG is trained on images of size 224×224 for classification, we use larger image resolutions of 336×336, which is possible since all our layers are convolutional. We use a mini-batch of size 12 as larger batches exceeded our GPU memory. We apply real time data augmentation, where in each mini-batch, the data is augmented (e.g., flips, rotations) and the mini-batch is randomly sampled such that at least two samples contain each of the class labels. The remaining four are randomly sampled. For our ISIC-ISBI entry, we did not use data augmentation nor over-sampling, and stopped training after only 5 epochs, as empirically we found longer training yielded segmentations less sensitive to the dermoscopic criteria. For our subsequent experiments, we show experiments with and without data augmentation/over-sampling, train for 100 epochs, and choose the model that achieves the lowest loss over our validation set.

7.3 Results and Discussions

We trained our network over 1,700 images from the ISIC-ISBI 2017 skin analysis challenge, and used 300 images to monitor the network's performance with different hyperparameters. The public leaderboard consisted of 150 images, with a separate private leaderboard of 600 images. While several metrics were evaluated, the winner of the challenge was determined by the highest averaged Area Under the Receiver Operator Characteristic curve (AUROC). Our approach achieved the highest averaged AUROC when compared to the other entries. The results over both the public validation and private test sets were fairly consistent. The results for ours and competing approaches over the private test set of 600 images are shown

in Table. 7.1. We composed Table 7.1 from the online submission system [77], which was evaluated over a controlled submission server and only made public after the competition.

Table 7.1: Official results over the ISIC-ISBI 2017 test dataset. Results are divided by challenge entry and type of dermoscopic criteria. The *average* row presents the results averaged over all the dermoscopic criteria. *ACC* represents accuracy, *AP* represents average precision, *SEN* represents sensitivity, and, *SPC* represents specificity.

Entry	Dermoscopic Criteria	ACC	AUROC	AP	SEN	SPC
Lee [107]	pigment network	0.915	0.828	0.487	0.736	0.921
	negative network	0.905	0.762	**0.321**	**0.618**	0.906
	milia-like cysts	0.843	0.837	**0.421**	**0.832**	0.843
	streaks	0.961	0.900	**0.422**	**0.839**	0.961
	average	0.906	0.832	**0.413**	0.649	0.907
Shen [107]	pigment network	0.909	0.835	0.491	0.756	0.914
	negative network	0.917	0.762	0.317	0.606	0.919
	milia-like cysts	0.852	**0.838**	0.418	0.824	0.852
	streaks	0.978	0.896	0.411	0.815	0.978
	average	0.914	0.833	0.409	**0.665**	0.915
ours	pigment network	**0.951**	**0.945**	**0.582**	**0.803**	**0.956**
	negative network	**0.982**	**0.869**	0.152	0.428	**0.984**
	milia-like cysts	**0.988**	0.807	0.078	0.303	**0.990**
	streaks	**0.997**	**0.960**	0.151	0.637	**0.997**
	average	**0.980**	**0.895**	0.241	0.542	**0.981**

7.3.1 Detecting Dermoscopic Criteria - Challenge Results

From Table 7.1, we observe the challenges and importance of choosing appropriate metrics when evaluating different methods. In addition to the metric of AUROC, accuracy, average precision, sensitivity, and specificity, were also evaluated. While AUROC was chosen as single metric to rank entries, and our approach achieved higher AUROC when compared to the other entries (ours 0.895 vs second place 0.833 [107]), the other entries outperform our approach on other metrics.

As the entry by Li and Shen [107] is a superpixel classification approach using a CNN, evaluated over the same dataset, we can compare superpixel classification with our semantic segmentation approach. In general, we see that our approach is less sensitive, but more specific when detecting dermoscopic criteria. Notably, for the *pigment network* dermoscopic criterion, we achieve the highest results across all metrics.

We show example results of the predicted and ground truth pixels for the different types of dermoscopic criteria in Fig. 7.4. This figure highlights the challenges of detecting dermoscopic criteria, as the visual cues for the various criteria are subtle and often not obvious to an untrained eye. We observe that *pigment network* and *streaks* often occur near the boundary of the lesion, while *negative network* can occur within the lesion. This

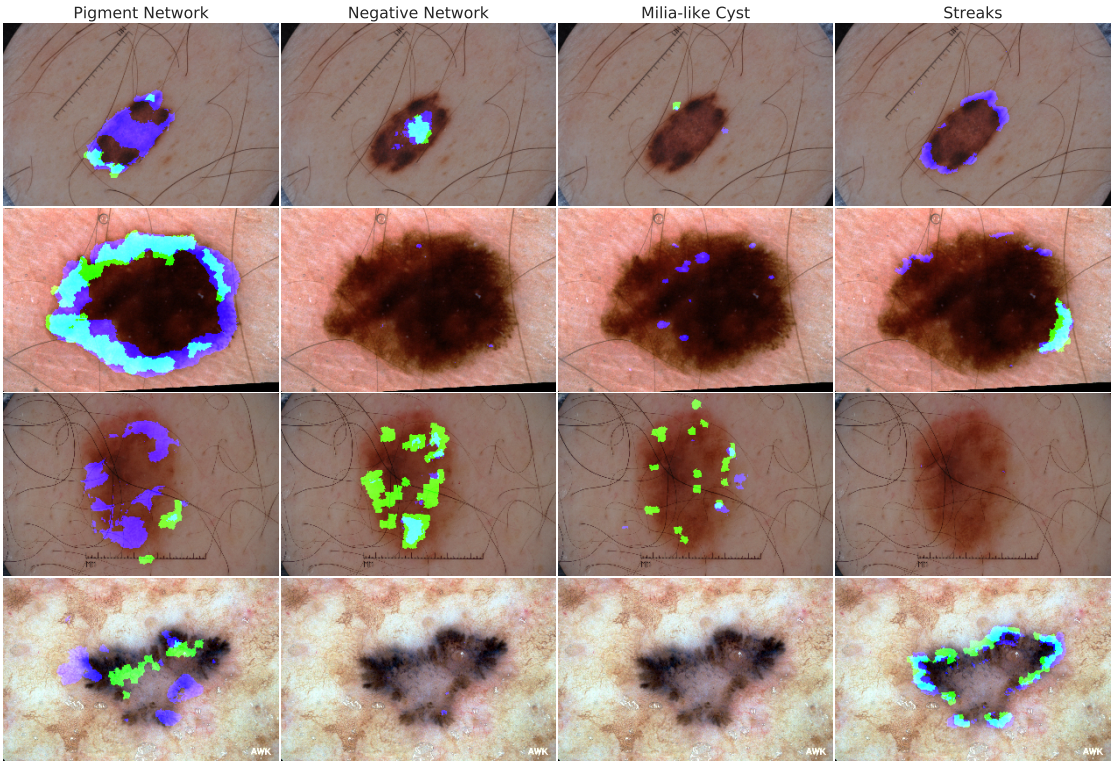


Figure 7.4: Dermoscopic criteria overlaid on the skin images. Each type of dermoscopic criterion (columns) is overlaid on four sample images from the test set (rows). *Green* pixels indicate ground truth. *Dark blue* pixels represent pixels predicted to have the specific feature. *Light blue* pixels indicate an overlap between predicted and ground truth.

illustrates how the context of the superpixel (i.e., information in surrounding pixels) is an important factor to consider when detecting dermoscopic criteria, and supports our approach to frame this task as segmentation problem, rather than classifying individual superpixels.

7.3.2 Detecting Dermoscopic Criteria - Simple Baselines

We show that two simple baseline approaches (Table 7.2 experiments *Lesion* and *Empty*) outperform existing methods when ranked using the metrics from Part 2 of the ISIC-ISBI 2016 [66] and 2017 [41] challenge. For the first baseline approach (Table 7.2 Exp. *Lesion*), we use a trained lesion segmentation model (described in Sec. 7.3.3) to label all pixels within a predicted lesion segmentation mask as positive incidences for all the dermoscopic criteria. Surprisingly, this simple approach achieves the highest averaged AUROC (used to rank Part 2 of the 2017 challenge [41]) and average precision score (used to rank Part 2A of the 2016 challenge [66]), outperforming existing methods (Table 7.2 Exp. *Lesion*). Although this approach scores high on the official benchmarks, classifying the entire lesion

Table 7.2: Two simple baselines experiments. *Lesion* indicates that the predicted lesion segmentation is used for all dermoscopic criteria predictions. *Empty* indicates that only background is predicted. *DC* is short for dermoscopic criteria. J_1 and J_{nan} represent the Jaccard Index with different values assigned to the empty set. Over the ISIC-ISBI 2017 test dataset, these simple baselines outperform existing methods when ranked using the challenge metrics, but not when ranked using the J_{nan} metric.

Exp.	DC	ACC	AUROC	AP	SEN	SPC	$\bar{J}_1$	$\bar{J}_{\text{nan}}$
Lesion	PN	0.832	0.913	0.528	0.962	0.827	0.167	0.167
	NN	0.807	0.916	0.502	0.992	0.806	0.012	0.012
	MC	0.805	0.884	0.421	0.915	0.805	0.016	0.016
	ST	0.803	0.894	0.380	0.960	0.803	0.001	0.001
	avg	0.812	**0.902**	0.458	**0.957**	0.810	0.049	0.049
Empty	PN	0.969	0.500	0.515	0.000	1.000	0.445	0.000
	NN	0.996	0.500	0.502	0.000	1.000	0.925	0.000
	MC	0.998	0.500	0.501	0.000	1.000	0.755	0.000
	ST	1.000	0.500	0.500	0.000	1.000	0.985	0.000
	avg	**0.991**	0.500	**0.505**	0.000	**1.000**	**0.777**	0.000
ours* (*ISIC entry*)	PN	0.951	0.944	0.585	0.806	0.956	0.319	0.217
	NN	0.982	0.870	0.159	0.427	0.984	0.339	0.021
	MC	0.988	0.809	0.075	0.294	0.990	0.225	0.031
	ST	0.997	0.963	0.154	0.605	0.997	0.532	0.007
	avg	0.980	0.896	0.243	0.533	0.982	0.354	**0.069**

*We report slight (≈1%) differences from the official results in Table 7.1.

as containing all types of dermoscopic criteria is not practically useful. In order to establish a metric that better captures the utility of the results, we propose to use a fuzzy Jaccard Index [44], defined as,

$$J(\hat{a}, a) = f_{\text{nan}}\left(\frac{\sum_i^n \min(\hat{a}_i, a_i)}{\sum_i^n \max(\hat{a}_i, a_i)}\right) \tag{7.8}$$

where the $\min(\cdot,\cdot)$ and $\max(\cdot,\cdot)$ functions compute a probabilistic intersection and union, respectively; $f_{\text{nan}}(x) = nan$ if the denominator is 0 else x; and nan is a sentinel indicating an undefined value. Given a test set of N predicted $\hat{M} \in \mathbb{R}^{N\times K\times W\times H}$ and ground truth M segmentations, computing the Jaccard Index over the entire set (i.e., $J(\hat{M}, M)$), will bias results towards more frequently occurring classes. Computing the Jaccard Index for each channel separately, $J_c(\hat{M}^{:,k,:,:}, M^{:,k,:,:})$ (this is how Part 2B [66] appeared to be ranked), will reduce the contribution of images with a relatively small proportion of positive pixels. In order to give higher weight to images with smaller dermoscopic criteria, for the k-th channel, we average over each image,

$$\bar{J}_1(\hat{M}^{:,k,:,:}, M^{:,k,:,:}) = \frac{1}{N}\sum_i^N f_1(J(\hat{M}^{i,k,:,:}, M^{i,k,:,:})) \tag{7.9}$$

where $\hat{M}^{:,k,:,:}$ are all N predictions for the k-th channel. An intuitive function that considers *nan* values, is to let $f_1(x) = 1$ if $x = nan$, else x, which returns a Jaccard Index of 1 when there are neither any positive predicted nor ground truth cases (i.e., the empty set). Using this measure, our proposed approach (Table 7.2 Exp. *ours*) scores considerably higher than the *Lesion* experiment, suggesting that $\bar{J}_1$ is a more informative metric than AUROC or the average precision score. However, in imbalanced datasets where many images contain no positive labels (Fig. 7.3), a classifier that predicts *only* background can achieve a high score. We empirically show that by predicting only background (Table 7.2 Exp. *Empty*), we achieve a higher Jaccard Index. Thus, we propose

$$\bar{J}_{\text{nan}}(\hat{M}^{:,k,:,:}, M^{:,k,:,:}) = \frac{\sum_i^N f_0(J(\hat{M}^{i,k,:,:}, M^{i,k,:,:}))}{\sum_i^N f_{01}(J(\hat{M}^{i,k,:,:}, M^{i,k,:,:}))} \tag{7.10}$$

where $f_0(x) = 0$ if $x = nan$, else x and $f_{01}(x) = 0$ if $x = nan$ else 1. This excludes all images where both the predicted and ground truth do not include any positive samples. $\bar{J}_{\text{nan}}$ penalizes a model that only assigns a background label (Exp. *Empty*), and our approach (Exp. *ours*) produces consistently higher $\bar{J}_{\text{nan}}$ scores than the *Lesion* experiment. We note that when computing the Jaccard Index, rather than using the predictions m directly, we use the superpixel probabilities (e.g., Fig. 7.1 *d,g*), i.e., use $\hat{z}_{ik}$ in Eq. 7.1, where $\hat{z}_{ik} = 0$ if $\hat{z}_{ik} < 0.5$ else $\hat{z}_{ik}$. This is done to remove false positive superpixels. Quantitative results showing of averaged improvements after thresholding and converting to superpixel segmentation are given in Table 7.5.

7.3.3 Lesion Segmentation

While not a focus of this thesis, we note that our entry for *Part 1: Lesion Segmentation Task* ranked sixth out of 21 entries based on the Jaccard distance (ours 0.752 vs first place 0.765 [177]). For our segmentation entry, we used nearly the same model and loss as described in this chapter. Notable differences include: images were resized 224×224; the original feature maps were used from the deeper layers; an additional convolutional layer after the concatenated layer was added; and, the model was trained for 12 epochs with a batch size of eight. Our competitive results over the segmentation challenge using only minor modifications suggests both lesion segmentation (Part 1) and dermoscopic clinical feature detection (Part 2) can be approached in similar ways. Fig. 7.5 shows examples where the contours of the ground truth and the predicted lesions are overlaid on the original lesion images. We sampled lesions that have a computed Jaccard Index around the range of the top performing methods (sampled between 0.736 and 0.782 Jaccard Index), to show the variability and subjectivity of the lesion borders in certain cases. Given the subjectivity observed in defining precise lesion borders, and the similarity between the top performing approach [177] and ours (only a 0.013 Jaccard Index difference), we can conclude that our segmentation approach is competitive with current state-of-the-art methods.

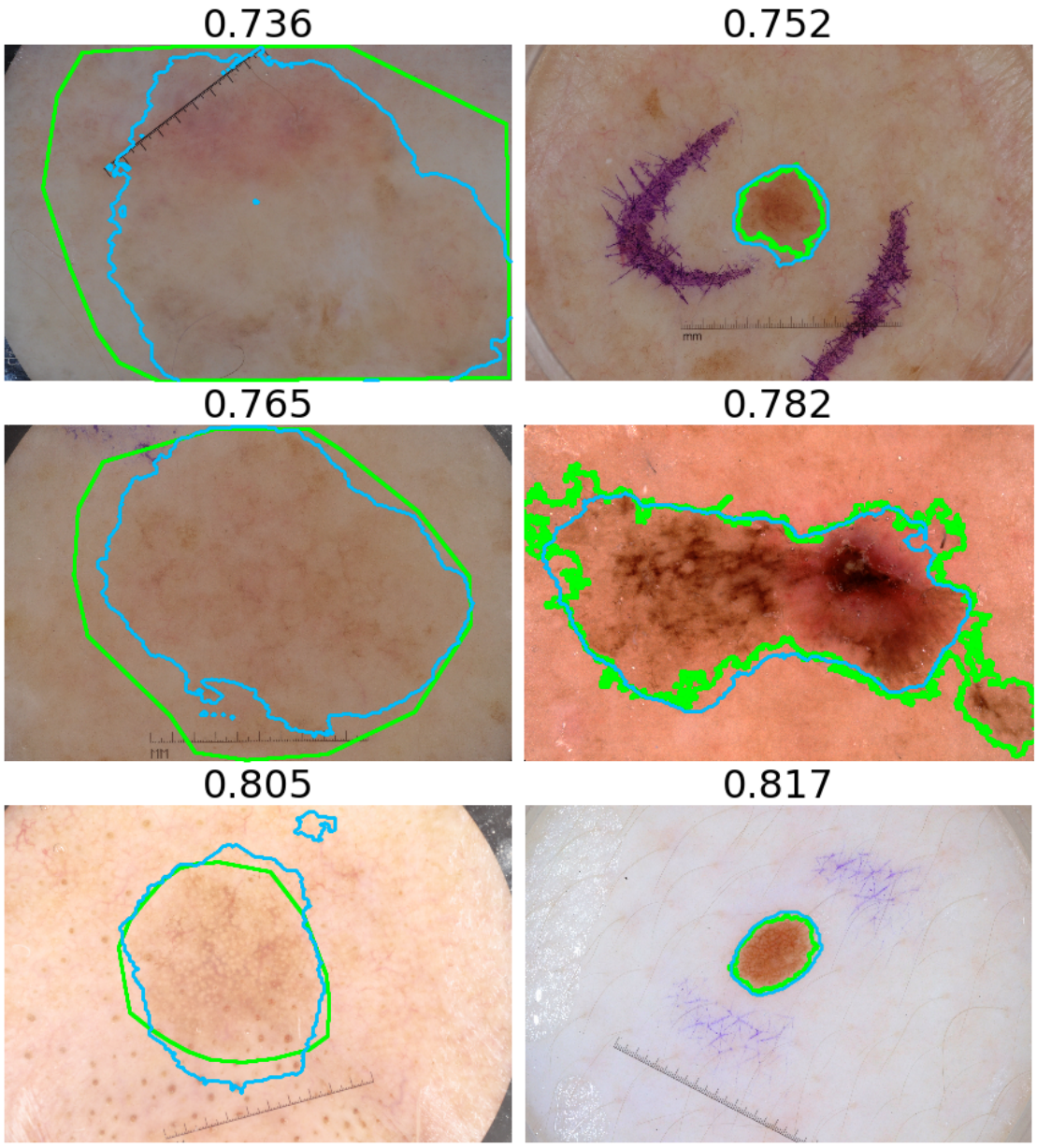


Figure 7.5: Example segmentation results where the *green line* indicates the ground truth contour, and the *blue line* represents our predicted lesion contour. The Jaccard Index between the predicted and ground truth lesion are displayed above each image. These cases illustrate where the exact lesion borders may be subjective. Note the variability in the ground truth borders (e.g., some have straight lines, while others are highly sensitive to intensity changes).

Table 7.3: Detailed results comparing loss functions. The *cross-entropy* loss is weighted to account for class imbalance. We display the ranking metrics, and note that while experiment *(a)* achieves the highest AUROC, we propose that the Jaccard Index $\bar{J}_{\text{nan}}$ better quantifies the performance of a model at the intended task.

Exp.	DC	AUROC	AP	J_{c}	$\bar{J}_{\text{nan}}$
	PN	0.963	0.578	0.299	0.209
(a)	NN	0.941	0.091	0.066	0.027
Cross-entropy	MC	0.948	0.077	0.037	0.023
class-weighted	ST	0.966	0.049	0.027	0.009
	avg	**0.955**	0.199	0.107	0.067
	PN	0.882	0.591	0.427	0.269
(b)	NN	0.502	0.008	0.000	0.000
Dice-F1	MC	0.500	0.001	0.000	0.000
volume-mini-batch	ST	0.500	0.000	0.000	0.000
Eq. 7.3	avg	0.596	0.150	0.107	0.067
	PN	0.938	0.591	0.380	0.232
(c)	NN	0.798	0.113	0.080	0.027
Dice-F1	MC	0.793	0.075	0.094	0.045
channel-image	ST	0.845	0.033	0.046	0.010
Eq. 7.6	avg	0.843	0.203	0.150	0.078
	PN	0.910	0.602	0.426	0.282
(d)	NN	0.645	0.134	0.079	0.056
Dice-F1	MC	0.737	0.103	0.126	0.051
channel-batch	ST	0.641	0.053	0.048	0.039
Eq. 7.7	avg	0.733	**0.223**	**0.170**	**0.107**

7.3.4 Comparing Losses and Model Variants

We compare the Dice-F_1 loss function with a weighted binary cross-entropy loss function, where we weight each pixel using median frequency balancing [52]. Using the weighted binary cross-entropy loss averaged over the four dermoscopic criteria as our loss function, the model converges to predicting *all* background labels (Table 7.4 - first row). Oversampling the minority class during data-augmentation improves results (Table 7.3*a*). While the resulting AUROC curve is higher than previously reported, the computed Jaccard Index is relatively low, indicating an over-segmentation similar to using the predicted lesion (Table 7.2 - *Lesion*).

Our subsequent experiments compare different mini-batch partitions when computing the Dice-F_1 score. When computing the Dice-F_1 score over the entire mini-batch over all labels (i.e., $D^* = D$ Eq. 7.5), only the larger *pigment network* class performs well (Table 7.3*b*). Averaging the loss over each mini-batch sample, over each label-channel (Eq. 7.6 $D^* = D^{B,K}$) further improved results (Table 7.3*c*). Computing the Dice-F_1 score over the entire channel within a mini-batch (Eq. 7.7 $D^* = D^K$), yields the top Jaccard Index (Table 7.3*d*).

Table 7.4: Experiments computing the loss over different mini-batch partitions and correcting for divide-by-zero errors. These results highlight the importance of choosing the appropriate mini-batch partition, and how subtle differences in correcting for divide-by-zero errors, or improper class weighting, can yield a model that converges to predicting all background values (denoted as $\bar{J}_{\text{nan}} = 0$).

Loss	Compute over		Class-augment	α	β	$\bar{J}_{\text{nan}}$
Cross-entropy	-	-	No	-	-	0.0
Cross-entropy	-	-	Yes	-	-	0.067
Dice-F1	Volume	Batch	Yes	0	1	0.067
Dice-F1	Channel	Image	Yes	1	1	0.0
Dice-F1	Channel	Image	Yes	0	1	0.078
Dice-F1	Channel	Batch	No	1	1	0.0
Dice-F1	Channel	Batch	No	0	1	0.083
Dice-F1	Channel	Batch	Yes	0	1	0.107

In Table 7.4, we show the model performance with setting α, β in Eq. 7.3 and through class oversampling during data augmentation. The cases where the model converges to predicting all background ($\bar{J}_{\text{nan}}$=0) indicates the challenges with infrequent class labels within imbalanced datasets.

Table 7.5 experiments with substituting VGG16 with more recent models: ResNet50 [74], and InceptionResNetV2 [152]. We find that changing the underlying model did not improve results. We suspect VGG is particularly well suited to this task since the first two convolutional layers of VGG16 maintain the original spatial dimensions of the input, producing high resolution feature maps that are directly considered in the output segmentation layer (in contrast ResNet50 reduces the spatial dimension in half after the first convolutional layer). As the dermoscopic criteria occupy only a fraction of the entire image, these high resolution feature maps may be necessary to detect subtle image cues.

Our final experiment replaces the concatenated skip connections with UNet [138] connections (Table 7.5). This did not improve the final result after thresholding and converting to superpixels. This may in part be due to the increased number of parameters that need to be learned to incorporate deeper feature maps. While these more recent models and modifications to the architecture did not improve results, we highlight that the Dice-F_1 loss function is not model specific, and other segmentation models may yield further improvements.

7.4 Conclusions

Our method approached the superpixel labelling task as a segmentation problem, used a CNN architecture that relied on interpolated and concatenated feature maps from the intermediate network layers, and minimized a negative multi-label Sørensen-Dice coefficient (F_1 score) computed across a partition of the mini-batch. We ranked first place in the ISIC-ISBI Part 2 Challenge, achieving the highest averaged area under the receiver operator

Table 7.5: Base models and segmentation connection types experiments. Using VGG as a base model with concatenated *skip* connections yielded slightly high averaged Jaccard Index results than other models and UNet type connections. This table also shows the results after using the direct prediction (*Direct*-$\bar{J}_{\text{nan}}$), after thesholding the predictions (*Thresh*-$\bar{J}_{\text{nan}}$), and converting the predictions to superpixels ($\bar{J}_{\text{nan}}$).

Base-model	Type	Direct-$\bar{J}_{\text{nan}}$	Thresh-$\bar{J}_{\text{nan}}$	$\bar{J}_{\text{nan}}$
InceptionResNetV2 [152]	Skip [112]	0.045	0.078	0.082
ResNet50 [74]	Skip	0.049	0.083	0.091
VGG [147]	UNet [138]	0.072	0.073	0.082
VGG	Skip	0.071	0.088	0.107

characteristic curve over both the public validation and private test-set leaderboard. For the specific dermoscopic criteria, we had the highest AUROC score for pigment network, negative network, and streaks. We demonstrated how simple baseline methods rank higher than existing approaches when using the current ranking metrics, and propose to use the averaged fuzzy Jaccard Index that ignores the values of the empty set. We highlight that the very low results reported using the averaged Jaccard Index from our top performing model (0.107), indicates significant room for improvement in this task, which is not as obvious when reporting the high (0.896) AUROC score. The ability to detect *pigment network* within dermoscopic images shows promise, although the low average precision and Jaccard Index indicates this task can be greatly improved. The low performance detecting other dermoscopic criteria remains an area for future research. Our competitive results over the Part 1 Segmentation challenge using nearly the same method, suggests both segmentation and clinical feature detection can be approached in similar ways. We hope the release of our code and trained model will serve as a baseline approach on which other groups can improve.

Chapter 8

Conclusions

8.1 Summary and Conclusions

This thesis proposed approaches to advance image-based machine diagnosis of skin lesions. Chapter 2 reviewed the literature on human and machine approaches to classify skin lesions from images and observed that several studies now report comparable human and machine classification accuracy (Section 2.3), with CNNs commonly being used in the top performing machine approaches (Section 2.2.4). Chapter 3 presented the first published work to evaluate whether the parameters of a CNN learned from non-skin images transfer well to multi-class (greater than two) lesion classification of clinical dermatological images. As these pretrained parameters worked well for classification, Chapter 4 represented dermoscopy and clinical skin images using the responses of a pretrained CNN, and proposed a visualization approach that retrieved a "path" of progressively similar skin images between two query images. While these pretrained parameters generalized to dermatological images, the parameters of the CNN used in Chapter 3 and Chapter 4 were not trained on the targeted dermatological domain. Chapter 5 proposed an end-to-end multi-resolution CNN architecture to classify skin lesions, where the CNN parameters are learned over multiple image resolutions.

In addition to directly diagnosing the skin lesion, Chapter 6 detected visual characteristics of the lesion that are commonly associated with a melanoma diagnosis. This chapter proposed a multi-modal CNN architecture and loss function designed for multiple tasks that was robust to missing data during inference. Finally, Chapter 7 proposed to detect (i.e., classify and localize) skin lesion characteristics associated with melanoma with a semantic segmentation CNN that was optimized to minimize a negative multi-label mini-batch-partitioned Dice score.

The CNN played a key component in all these works, where each proposed work modified the traditional CNN architecture to consider the types of input data and to suit a variety of clinical tasks. The CNN primarily served as the role of a strong feature extractor, in which the high dimensional input data (e.g., images) was transformed to a lower dimensional feature vector through a learned transformation. These chapters showed a progression starting

with the CNN used as a fixed feature extractor for classification (Chapter 3) and image retrieval (Chapter 4). This was followed by trainable architectures (Chapter 5) capable of classifying (Chapter 6) and detecting (Chapter 7) dermoscopic criteria. The building blocks within the CNN provided a single optimization framework, in which a variety of clinical tasks were defined and optimized for.

8.2 Future Directions

This thesis has a large focus on dermoscopy images and melanoma diagnosis, which covers a relatively narrow scope of the entire dermatology field. While this thesis proposed approaches which considered clinical skin images spanning 10 classes of skin lesions, the general problem of diagnosing thousands of types of skin diseases from non-standardized clinical skin images (e.g., images captured via a mobile phone) offers additional challenges. This section asks open questions that may serve future collaborating computing and dermatology researchers, especially those looking to transition a promising machine approach to clinical practice.

8.2.1 How to measure the clinical utility of a model?

The choice of a metric to measure performance is non-trivial and should be considered in the context of clinical utility. For example, Section 7.3.2 showed how using lesion segmentations as a predicted dermoscopic criteria mask produces an AUROC score of 0.9. By this metric, we may mistakenly conclude this model has high clinical utility. However, the segmented lesion is arguably not clinically useful for detecting dermoscopic criteria (since the lesion appears in every image, but the dermoscopic criteria do not). When measured with a modified Jaccard Index, the reported score of less than 0.1 indicates automated approaches have significant room for improvement. For skin lesion classification of multiple types of skin lesions, global accuracy does not well capture the performance of infrequent conditions, which are frequently the class of interest (e.g., melanoma). Average sensitivity weighs all types of diseases equally, but does not consider the clinical penalty of misdiagnoses (e.g., a low clinical penalty when misclassifying types of benign lesions, but a high clinical penalty of misclassifying a malignant lesion as benign). An improved metric may incorporate the clinical penalty for misdiagnosis of certain disease types, and could likely be directly optimized for within a machine learning approach. A recent dermatologist led study, measured the correctly predicted clinical management approach [148], which may also be a promising direction for establishing clinically relevant metrics.

8.2.2 What is the diagnostic performance of human dermatologists on a large scale multi-class clinical image dataset?

While studies have been done to measure the diagnostic performance of human dermatologists on images (Section 2.3), the general diagnostic performance of dermatologists across many skin conditions is unclear. A better understanding of human performance would help computer vision researchers estimate a reasonable target for the performance of vision-based machine models. There may be an upper-bound to classification performance due to insufficient information within images alone.

An ideal study would have multiple human dermatologists independently classify a variety of skin diseases from publicly available images, where they output a list of ranked diagnoses based on a large set of possible conditions. This is similar to the recent study by Brinker et al. [26], which will publicly release the performance of approximately 150 dermatologists on the binary task of melanoma classification from 100 publicly available clinical and dermoscopy images. As the ideal metric to measure performance is debatable, a study that released the ranked predicted diagnoses of each dermatologist would allow new metrics to be computed over time and may give a clearer understanding of the variability among dermatologists.

Such a study may also address the earlier question of how to measure the clinical utility of the model (Sec. 8.2.1), where a visual variant of a Turing Test [163] could be performed. If a machine model outputs predictions that are indistinguishable from human dermatologists over a large varied dataset, we may conclude that the predictions made by the machine model have the clinical utility of a human dermatologist. We highlight that although machines and humans may achieve comparable diagnostic performance, other clinical factors, such as offering support to patients and explaining the diagnosis and treatment options, are important considerations not included in this measure.

8.2.3 How to incorporate information beyond photographs and what information to gather?

As we reach the upper bound of image-based diagnosis, incorporating modalities such as video, depth, or other non-visual information (e.g., patient history, age, sex) may improve diagnostic performance. Although, in general, the reported performance of image-only diagnosis is similar to the performance when non-visual patient information is included (Fig. 2.7), datasets that include more types of skin disease conditions may benefit from this type of additional information. For example, common dermatological diseases differ by patient age [101], which could be encoded as prior knowledge. Further challenges remain such as how to determine what clinically relevant patient information should be captured (e.g., Is the lesion itchy? Was there trauma to the location?) and how to encode this data. While the multi-modal model and loss functions proposed in this thesis (Fig. 6.1) is an initial attempt to incorporate differing types of dermatological data, this could be extended to ad-

dress missing training data and reduce the large number of combinations of loss functions required for multiple modalities (Eq. 6.3).

8.2.4 How to diagnose at an appropriate level of abstraction?

Each type of skin condition can be labelled at different levels of abstraction, from a *general* label that encompasses multiple types of skin conditions (e.g., *benign nevus* describes several types of nevi) to a *specific* label that describes distinct lesion characteristics (e.g., *blue nevus* describes a benign nevi of a bluish colour). Choosing a label to output to the user (i.e., a *diagnosis*) at an appropriate level of abstraction is challenging. A general diagnosis may be accurate, but may cover too many conditions to recommend appropriate treatment. A specific diagnosis may offer more information about the condition, but be less accurate. Further, a specific diagnosis may have the same disease management as a more general diagnosis, and thus offer minimal benefit to the patient. This is further complicated as there exists different terminologies and structures to organize the relationships among the terminologies, such as the hierarchical tree of ICD-10 [171] and the directed acyclic graph of Derm0 [59]. This thesis predicted the most specific label available for each skin condition (with the exception of Chapter 6, which aggregated conditions due to a limited number of images). Future research may leverage the relationships among conditions and consider the clinical management and model performance to explore the trade-offs with diagnosing at different levels of abstraction.

8.2.5 How to make an interpretable machine diagnosis?

We do not know if a "black box" machine diagnosis, in which a diagnosis is given but no explanation of the decision is provided, will be adopted by the public. Understanding *why* a diagnosis is made may help patients and health care providers accept a diagnosis. As discussed by Barata et al. [17], diagnosis systems that base decisions on medically interpretable features may allow dermatologists to better understand and validate a machine diagnosis.

Chapter 4 and Chapter 6 explored retrieving images from a known reference dataset of labelled skin images as a way of providing a more interpretable diagnosis through showing examples. More elaborate visualization approaches, such as the path of similar skin images proposed in Chapter 4, may play an important role in understanding results through guiding the user to relevant images of known conditions. Chapter 6 explored visualizing the image regions that influenced the CNN's predictions as a way to understand the decision. Chapter 6 classified and Chapter 7 localized dermoscopic criteria known to be associated with melanoma, which gave insights into the visual properties within the lesion that relate to the diagnosis. Extending this approach to other conditions may give insights into the underlying visual properties that indicate a particular diagnosis; however, this will require significant clinical effort to gather and label the relevant dermoscopic criteria for each type of skin condition.

Bibliography

[1] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems. `http://tensorflow.org/`, 2015. [Accessed May 22, 2019].

[2] Q. Abbas, I. F. Garcia, M. E. Celebi, and W. Ahmad. A feature-preserving hair removal algorithm for dermoscopy images. *Skin Research and Technology*, 19(1):e27–36, 2013.

[3] American Cancer Society. Melanoma skin cancer. `https://www.cancer.org/cancer/melanoma-skin-cancer.html`. [Accessed May 22, 2019].

[4] Z. Apalla, D. Nashan, R. B. Weller, and X. Castellsagué. Skin cancer: Epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. *Dermatology and Therapy*, 7:5–19, 2017.

[5] G. Argenziano et al. Interactive atlas of dermoscopy: A tutorial (Book and CD-ROM). *EDRA Medical Publishing & New Media*, 2000. `http://www.dermoscopy.org/atlas/order_cd.asp`. [Accessed May 22, 2019].

[6] G. Argenziano et al. Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the internet. *Journal of the American Academy of Dermatology*, 48(5):679–693, 2003.

[7] G. Argenziano et al. Seven-point checklist of dermoscopy revisited. *British Journal of Dermatology*, 164(4):785–790, 2011.

[8] G. Argenziano et al. Total body skin examination for skin cancer screening in patients with focused symptoms. *Journal of the American Academy of Dermatology*, 66(2):212–219, 2012.

[9] G. Argenziano, G. Fabbrocini, P. Carli, D. G. Vincenzo, E. Sammarco, and M. Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of Dermatology*, 134(12):1563–1570, 1998.

[10] R. Ashton, B. Leppard, and H. Cooper. *Differential Diagnosis in Dermatology*. CRC Press. Taylor & Francis Group, Boca Raton, fourth edition, 2014.

[11] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees. Non-melanoma skin lesion classification using colour image data in a hierarchical K-NN classifier. In *IEEE International Symposium on Biomedical Imaging*, pages 358–361, 2012.

[12] 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 M. E. Celebi and G. Schaefer, editors, *Color Medical Image Analysis*, volume 6, pages 63–86. Springer Netherlands, 2013.

[13] L. Ballerini, X. Li, R. Fisher, B. Aldridge, and J. Rees. Content-based image retrieval of skin lesions by evolutionary feature synthesis. In *EvoApplications*, pages 312–319. Springer-Verlag, 2010.

[14] S. Bao and A. C. S. Chung. Multi-scale structured CNN with label consistency for brain MR image segmentation. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 113–117, 2016.

[15] C. Barata, E. M. Celebi, and J. S. Marques. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2018.

[16] C. Barata, M. E. Celebi, and J. S. Marques. Improving dermoscopy image classification using color constancy. *IEEE Journal of Biomedical and Health Informatics*, 19(3):1146–1152, 2015.

[17] C. Barata, M. Emre Celebi, and J. S. Marques. Development of a clinically oriented system for melanoma diagnosis. *Pattern Recognition*, 69:270–285, 2017.

[18] C. Barata, J. S. Marques, and J. Rozeira. A system for the detection of pigment network in dermoscopy images using directional filters. *IEEE Transactions on Biomedical Engineering*, 59(10):2744–2754, oct 2012.

[19] F. C. Beddingfield. The melanoma epidemic: Res ipsa loquitur. *The Oncologist*, 8(5):459–465, 2003.

[20] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*, 34(5):321–359, 2007.

[21] A. BenTaieb, J. Kawahara, and G. Hamarneh. Multi-loss convolutional networks for gland analysis in microscopy. In *IEEE International Symposium on Biomedical Imaging*, pages 642–645, 2016.

[22] D. R. Bickers, H. W. Lim, D. Margolis, M. a. Weinstock, C. Goodman, E. Faulkner, C. Gould, E. Gemmen, and T. Dall. The burden of skin diseases: 2004. A joint project of the American Academy of Dermatology Association and the Society for Investigative Dermatology. *Journal of the American Academy of Dermatology*, 55(3):490–500, 2006.

[23] Y. Boykov and G. Funka-Lea. Graph cuts and efficient N-D image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.

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

[25] T. J. Brinker. The melanoma classification benchmark. `https://skinclass.de/mclass/`. [Accessed May 22, 2019].

[26] T. J. Brinker, A. Hekler, A. Hauschild, C. Berking, B. Schilling, A. H. Enk, S. Haferkamp, A. Karoglan, C. von Kalle, M. Weichenthal, E. Sattler, D. Schadendorf, M. R. Gaiser, J. Klode, and J. S. Utikal. Comparing artificial intelligence algorithms to 157 German dermatologists: The melanoma classification benchmark. *European Journal of Cancer*, 111:30–37, 2019.

[27] T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk, and C. von Kalle. Skin cancer classification using convolutional neural networks: Systematic review. *Journal of Medical Internet Research*, 20(10):e11936, 2018.

[28] C. J. Brown, J. Kawahara, and G. Hamarneh. Connectome priors in deep neural networks to predict autism. In *IEEE International Symposium on Biomedical Imaging*, pages 110–113, Washington, D.C., 2018.

[29] K. Bunte, M. Biehl, M. F. Jonkman, and N. Petkov. Learning effective color features for content based image retrieval in dermatology. *Pattern Recognition*, 44(9):1892–1902, 2011.

[30] P. Carli, V. De Giorgi, G. Argenziano, D. Palli, and B. Giannotti. Pre-operative diagnosis of pigmented skin lesions: In vivo dermoscopy performs better than dermoscopy on photographic images. *Journal of the European Academy of Dermatology and Venereology*, 16(4):339–346, 2002.

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

[32] M. E. Celebi and Y. A. Aslandogan. Content-based image retrieval incorporating models of human perception. In *IEEE International Conference on Information Technology: Coding and Computing*, pages 241–245, 2004.

[33] M. E. Celebi, N. Codella, and A. Halpern. Dermoscopy image analysis: Overview and future directions. *IEEE Journal of Biomedical and Health Informatics*, 23(2):474–478, 2019.

[34] M. E. Celebi, H. Iyatomi, W. V. Stoecker, R. H. Moss, H. S. Rabinovitz, G. Argenziano, and H. P. Soyer. Automatic detection of blue-white veil and related structures in dermoscopy images. *Computerized Medical Imaging and Graphics*, 32(8):670–677, 2008.

[35] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss. A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6):362–373, 2007.

[36] M. E. Celebi, Q. Wen, H. Iyatomi, K. Shimizu, H. Zhou, and G. Schaefer. A state-of-the-art survey on lesion border detection in dermoscopy images. In M. E. Celebi,

T. Mendonça, and J. S. Marques, editors, *Dermoscopy Image Analysis*, chapter 4, pages 97–130. CRC Press. Taylor & Francis Group, 1st edition, 2015.

[37] F. Chollet et al. Keras. `https://keras.io/`, 2015. [Accessed May 22, 2019].

[38] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similiarty metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 349–356, 2005.

[39] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *ArXiv e-prints:1902.03368*, pages 1–12, 2019.

[40] N. C. F. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning In Medical Imaging*, volume 9352, pages 118–126, 2015.

[41] 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, A. Halpern, and C. Park. 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.

[42] N. C. F. Codella, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, J. R. Smith, and U. States. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4–5):1–15, 2017.

[43] D. E. Cohen, O. P. Sangueza, E. Fass, and M. J. Stiller. In vivo cutaneous surface microscopy: Revised nomenclature. *International Journal of Dermatology*, 32(4):257–258, 1993.

[44] W. R. Crum, O. Camara, and D. L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

[45] R. C. da Silva, M. Bevilaqua, and J. de Morais Silva. Stains. In R. Bonamigo and S. Dornelles, editors, *Dermatology in Public Health Environments*, chapter 58, pages 1199–1227. Springer, 2018.

[46] S. Demyanov, R. Chakravorty, Z. Ge, S. Bozorgtabar, M. Pablo, A. Bowling, and R. Garnavi. Tree-loss function for training neural networks on weakly-labelled datasets. In *IEEE International Symposium on Biomedical Imaging*, pages 287–291, 2017.

[47] Dermatology database used in MED-NODE. `http://www.cs.rug.nl/~imaging/databases/melanoma_naevi/`. [Accessed May 22, 2019].

[48] DermQuest. `https://www.derm101.com/dermquest/`. [Accessed May 22, 2019].

[49] T. L. Diepgen. Dermatology information system. `http://www.dermis.net`. [Accessed May 22, 2019].

[50] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, volume 32, pages 647–655, 2014.

[51] K. Duffy and D. Grossman. The dysplastic nevus: From historical perspective to management in the modern era: Part I. Historical, histologic, and clinical aspects. *Journal of the American Academy of Dermatology*, 67(1):1–27, 2012.

[52] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[53] J. G. Elmore, R. L. Barnhill, D. E. Elder, G. M. Longton, M. S. Pepe, L. M. Reisch, P. A. Carney, L. J. Titus, H. D. Nelson, T. Onega, A. N. Tosteson, M. A. Weinstock, S. R. Knezevich, and M. W. Piepkorn. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: Observer accuracy and reproducibility study. *BMJ*, 357:1–11, 2017.

[54] 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):115–118, 2017.

[55] G. Fabbrocini, V. De Vita, S. Cacciapuoti, G. Di Leo, C. Liguori, A. Paolillo, A. Pietrosanto, and P. Sommella. Automatic diagnosis of melanoma based on the 7-point checklist. In J. Scharcanski and M. E. Celebi, editors, *Computer Vision Techniques for the Diagnosis of Skin Cancer*, pages 71–107. Springer-Verlag Berlin Heidelberg, 2014.

[56] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[57] P. M. Ferreira. PH^2 database. `https://www.fc.up.pt/addi/`, 2012. [Accessed May 22, 2019].

[58] L. K. Ferris, J. A. Harkes, B. Gilbert, D. G. Winger, K. Golubets, O. Akilov, and M. Satyanarayanan. Computer-aided classification of melanocytic lesions using dermoscopic images. *Journal of the American Academy of Dermatology*, 73(5):769–776, 2015.

[59] H. M. Fisher, R. Hoehndorf, B. S. Bazelato, S. S. Dadras, L. E. King, G. V. Gkoutos, J. P. Sundberg, and P. N. Schofield. DermO; An ontology for the description of dermatologic disease. *Journal of Biomedical Semantics*, 7(38):1–9, 2016.

[60] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

[61] Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, and M. Fujimoto. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *British Journal of Dermatology*, 180(2):373–381, 2019.

[62] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, 2016.

[63] Z. Ge, S. Demyanov, B. Bozorgtabar, M. Abedini, Rajib Chakravorty, A. Bowling, and R. Garnavi. Exploiting local and generic features for accurate skin lesions classification using clinical and dermoscopy imaging. In *IEEE International Symposium on Biomedical Imaging*, pages 986–990, 2017.

[64] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *Medical Image Computing and Computer Assisted Intervention*, volume 10435, pages 250–258. Springer, 2017.

[65] I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications*, 42(19):6578–6585, 2015.

[66] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *ArXiv e-prints:1605.01397*, pages 1–5, 2016.

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

[68] H. A. Haenssle, B. Korpas, C. Hansen-Hagge, T. Buhl, K. M. Kaune, A. Rosenberger, U. Krueger, M. P. Schön, and S. Emmert. Seven-point checklist for dermatoscopy: Performance during 10 years of prospective surveillance of patients at increased melanoma risk. *Journal of the American Academy of Dermatology*, 62(5):785–793, 2010.

[69] G. Hamarneh, A. Amir-Khalili, M. Nosrati, I. Figueroa, J. Kawahara, O. Al-Alao, J.-M. Peyrat, J. Abi-Nahed, A. Al-Ansari, and R. Abugharbieh. Towards multi-modal image-guided tumour identification in robot-assisted partial nephrectomy. In *IEEE Middle East Conference on Biomedical Engineering*, pages 159–162, 2014.

[70] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538, 2018.

[71] S. S. Han, G. H. Park, W. Lim, M. S. Kim, J. I. Na, I. Park, and S. E. Chang. Deep neural networks show an equivalent and often superior performance to dermatologists

in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS ONE*, 13(1):1–14, 2018.

[72] R. J. Hay, N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, C. Michaud, C. J L Murray, and M. Naghavi. The global burden of skin disease in 2010: An analysis of the prevalence and impact of skin conditions. *The Journal of Investigative Dermatology*, 134(6):1527–1534, 2014.

[73] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *European Conference on Computer Vision*, volume 8691, pages 346–361. Springer, 2014.

[74] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[75] C. Hegde, A. C. Sankaranarayanan, and R. G. Baraniuk. Learning manifolds in the wild. Technical report, Rice University, 2012.

[76] K. Hogan, J. Cullan, V. Patel, A. Rajpara, and D. Aires. Overcalling a teledermatology selfie: A new twist in a growing field. *Dermatology Online Journal*, 21(6):0–4, 2015.

[77] International Skin Imaging Collaboration. Part 2: Lesion dermoscopic feature extraction - Phase 3: Final test submission. `https://challenge.kitware.com/#phase/584b0afacad3a51cc66c8e2e`, 2017. [Accessed May 22, 2019].

[78] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *International Conference on Machine Learning*, volume 37, pages 448–456. PMLR, 2015.

[79] S. Izadi, Z. Mirikharaji, J. Kawahara, and G. Hamarneh. Generative adversarial networks to segment skin lesions. In *IEEE International Symposium on Biomedical Imaging*, pages 881–884, Washington, D.C., 2018.

[80] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[81] H. Jia, G. Wu, Q. Wang, Y. Wang, M. Kim, and D. Shen. Directed graph based image registration. In K. Suzuki, F. Wang, D. Shen, and P. Yan, editors, *International Workshop on Machine Learning In Medical Imaging*, volume 7009, pages 175–183. Springer, 2011.

[82] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678, 2014.

[83] R. H. Johr. Interactive CD of dermascopy. *Archives of Dermatology*, 137(6):831–832, 2001.

[84] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker. Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. In *MICCAI Workshop on Ischemic Stroke Lesion Segmentation*, pages 13–16, 2015.

[85] J. Kawahara, A. BenTaieb, and G. Hamarneh. Deep features to classify skin lesions. In *IEEE International Symposium on Biomedical Imaging*, pages 1397–1400, 2016.

[86] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146(Feb):1038–1049, 2017.

[87] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. 7-point criteria evaluation database. `http://derm.cs.sfu.ca/`, 2018. [Accessed May 22, 2019].

[88] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

[89] J. Kawahara and G. Hamarneh. Image content-based navigation of skin conditions. In *World Congress of Dermatology*, page 1 [abstract], 2015.

[90] J. Kawahara and G. Hamarneh. Multi-resolution-tract cnn with hybrid pretrained and skin-lesion trained layers. In L. Wang, E. Adeli, Q. Wang, Y. Shi, and H. Suk, editors, *International Workshop on Machine Learning In Medical Imaging*, pages 164–171. Springer, 2016.

[91] J. Kawahara and G. Hamarneh. Fully convolutional neural networks to detect clinical dermoscopic features. *IEEE Journal of Biomedical and Health Informatics*, 23(2):578–585, 2019.

[92] J. Kawahara, K. P. Moriarty, and G. Hamarneh. Graph geodesics to find progressively similar skin lesion images. In *International Workshop on Graphs in Biomedical Image Analysis*, volume 10551, pages 31–41. LNCS, 2017.

[93] J. Kawahara, J. M. Peyrat, J. Abinahed, O. Al-Alao, A. Al-Ansari, R. Abugharbieh, and G. Hamarneh. Automatic labelling of tumourous frames in free-hand laparoscopic ultrasound video. In P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 8674 LNCS Part II, pages 676–683. Springer International Publishing, 2014.

[94] A. B. Kimball and J. S. Resneck. The US dermatology workforce: A specialty remains in shortage. *Journal of the American Academy of Dermatology*, 59(5):741–745, 2008.

[95] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15, 2015.

[96] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder. Diagnostic accuracy of dermoscopy. *Lancet Oncology*, 3(3):159–165, 2002.

[97] M. Klingemann and S. Doury. X degrees of separation. `https://artsexperiments.withgoogle.com/xdegrees/`, 2016. [Accessed May 22, 2019].

[98] G. Kogan. Shortest path between images. `https://github.com/ml4a/ml4a-guides/blob/master/notebooks/image-path.ipynb`, 2017. [Accessed May 22, 2019].

[99] K. Korotkov and R. Garcia. Computerized analysis of pigmented skin lesions: A review. *Artificial Intelligence in Medicine*, 56(2):69–90, 2012.

[100] 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.

[101] E. T. Landis, S. A. Davis, A. Taheri, and S. R. Feldman. Top dermatologic diagnoses by age. *Dermatology Online Journal*, 20(4):1–6, 2014.

[102] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[103] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[104] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *International Conference on Artificial Intelligence and Statistics*, volume 38, pages 562–570, 2015.

[105] C. D. Leo, V. Bevilacqua, L. Ballerini, R. Fisher, B. Aldridge, and J. Rees. Hierarchical classification of ten skin lesion classes. In *SICSA Dundee Medical Image Analysis Workshop*, page 1, 2015.

[106] X. Li, B. Aldridge, L. Ballerini, R. Fisher, and J. Rees. Depth data improves skin lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention*, volume 5762, pages 1100–1107, 2009.

[107] Y. Li and L. Shen. Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. *Sensors*, 18(556):1–16, 2018.

[108] H. W. Lim, S. A. B. Collins, J. S. Resneck, J. L. Bolognia, J. A. Hodge, T. A. Rohrer, M. J. Van Beek, D. J. Margolis, A. J. Sober, M. A. Weinstock, D. R. Nerenz, W. Smith Begolka, and J. V. Moyano. The burden of skin disease in the United States. *Journal of the American Academy of Dermatology*, 76(5):958–972.e2, 2017.

[109] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations*, 2014.

[110] P. A. Lio and P. Nghiem. Interactive atlas of dermoscopy. *Journal of the American Academy of Dermatology*, 50(5):807–808, 2004.

[111] A. Lomas, J. Leonardi-Bee, and F. Bath-Hextall. A systematic review of worldwide incidence of nonmelanoma skin cancer. *British Journal of Dermatology*, 166(5):1069–1080, 2012.

[112] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[113] A. R. Lopez, X. Giro-i Nieto, J. Burdick, and O. Marques. Skin lesion classification from dermoscopic images using deep learning techniques. In *IASTED International Conference Biomedical Engineering*, pages 49–54, 2017.

[114] A. Madooei and M. S. Drew. Detecting specular highlights in dermatological images. In *IEEE International Conference on Image Processing*, pages 4357–4360, 2015.

[115] A. Madooei and M. S. Drew. Incorporating colour information for computer-aided diagnosis of melanoma from dermoscopy images: A retrospective survey and critical analysis. *International Journal of Biomedical Imaging*, Article ID 4868305:1–18, 2016.

[116] A. Madooei, M. S. Drew, M. Sadeghi, and M. S. Atkins. Automatic detection of blue-white veil by discrete colour matching in dermoscopy images. In *Medical Image Computing and Computer Assisted Intervention*, volume 8151, pages 453–460. Springer, 2013.

[117] J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[118] M. A. Marchetti, N. C. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M. E. Celebi, J. L. DeFazio, N. Jaimes, A. A. Marghoob, E. Quigley, A. Scope, O. Yélamos, and A. C. Halpern. 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):270–278, 2017.

[119] S. N. Markovic, L. A. Erickson, R. D. Rao, et al. Malignant melanoma in the 21st Century, Part 1: Epidemiology, risk factors, screening, prevention, and diagnosis. *Mayo Clinic Proceedings*, 82(3):364–380, 2007.

[120] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira. PH^2 - A dermoscopic image database for research and benchmarking. In *IEEE Engineering in Medicine and Biology Society*, pages 5437–5440, 2013.

[121] T. F. Mendonça, P. M. Ferreira, A. R. S. Marçal, C. Barata, J. S. Marques, J. Rocha, and J. Rozeira. Ph^2 A public database for the analysis of dermoscopic images. In M. E. Celebi, T. Mendonça, and J. S. Marques, editors, *Dermoscopy Image Analysis*, chapter 13, pages 419–439. CRC Press. Taylor & Francis Group, 1st edition, 2015.

[122] 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.

[123] F. Milletari, N. Navab, and S.-A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Fourth Int. Conf. on 3D Vis.*, pages 565–571, 2016.

[124] Z. Mirikharaji, S. Izadi, J. Kawahara, and G. Hamarneh. Deep auto-context fully convolutional neural network for skin lesion segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 877–880, Washington, D.C., 2018.

[125] H. Mirzaalian, T. K. Lee, and G. Hamarneh. Learning features for streak detection in dermoscopic color images using localized radial flux of principal intensity curvature. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 97–101, 2012.

[126] H. Mirzaalian, T. K. Lee, and G. Hamarneh. Hair enhancement in dermoscopic images using dual-channel quaternion tubularness filters and MRF-based multilabel optimization. *IEEE Transactions on Image Processing*, 23(12):5486–5496, 2014.

[127] N. K. Mishra and M. E. Celebi. An overview of melanoma detection in dermoscopy images using image processing and machine learning. *ArXiv e-prints:1601.07843*, pages 1–15, 2016.

[128] G. Monheit, A. B. Cognetta, L. Ferris, H. Rabinovitz, K. Gross, M. Martini, J. M. Grichnik, M. Mihm, V. G. Prieto, P. Googe, R. King, A. Toledano, N. Kabelev, M. Wojton, and D. Gutkowicz-Krusin. The performance of MelaFind: A prospective multicenter study. *Archives of Dermatology*, 147(2):188–194, 2011.

[129] 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.

[130] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696, 2011.

[131] R. Nock and F. Nielsen. Statistical region merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1452–1458, 2004.

[132] E. Okur and M. Turkan. A survey on automated melanoma detection. *Engineering Applications of Artificial Intelligence*, 73(April):50–67, 2018.

[133] J. Pastor-Pellicer, F. Zamora-Martínez, S. España-Boquera, and M. J. Castro-Bleda. F-measure as the error function to train neural networks. In *International Work-Conference on Artificial Neural Networks*, volume 7902, pages 376–384. Springer, 2013.

[134] S. Pathan, K. G. Prabhu, and P. C. Siddalingaswamy. Techniques and algorithms for computer aided diagnosis of pigmented skin lesions - A review. *Biomedical Signal Processing and Control*, 39:237–262, 2018.

[135] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In D. Stoyanov et al., editors, *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, volume 11041, pages 303–311. Springer, 2018.

[136] O. Razeghi and G. Qiu. 2309 skin conditions and crowd-sourced high-level knowledge dataset for building a computer aided diagnosis system. In *IEEE International Symposium on Biomedical Imaging*, pages 61–64, 2014.

[137] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[138] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, volume 9351, pages 234–241, 2015.

[139] M.-L. Rübsam, M. Esch, E. Baum, and S. Bösner. Diagnosing skin disease in primary care: A qualitative study of GPs' approaches. *Family Practice*, 32(5):591–595, 2015.

[140] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[141] M. Sadeghi, T. K. Lee, D. McLean, H. Lui, and M. S. Atkins. Detection and analysis of irregular streaks in dermoscopic images of skin lesions. *IEEE Transactions on Medical Imaging*, 32(5):849–861, 2013.

[142] J. K. Schofield, D. Fleming, D. Grindlay, and H. Williams. Skin conditions are the commonest new reason people present to general practitioners in England and Wales. *British Journal of Dermatology*, 165:1044–1050, 2011.

[143] K. Sellheyer and W. F. Bergfeld. A retrospective biopsy study of the clinical diagnostic accuracy of common skin diseases by different specialties compared with dermatology. *Journal of the American Academy of Dermatology*, 52(5):823–830, 2005.

[144] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. *International Conference on Learning Representations*, pages 1–16, 2014.

[145] K. Shimizu, H. Iyatomi, M. E. Celebi, K.-A. Norton, and M. Tanaka. Four-class classification of skin lesions with task decomposition strategy. *IEEE Transactions on Biomedical Engineering*, 62(1):274–283, 2015.

[146] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[147] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, pages 1–14, 2015.

[148] C. Sinz, P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J. Y. Gourhant, J. Kreusch, A. Lallas, J. Lapins, A. A. Marghoob, S. W. Menzies, J. Paoli, H. S. Rabinovitz, C. Rinner, A. Scope, H. P. Soyer, L. Thomas, I. Zalaudek, and H. Kittler. Accuracy of dermatoscopy for the diagnosis of non-pigmented cancers of the skin. *Journal of the American Academy of Dermatology*, 77(6):1100–1109, 2017.

[149] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *International Workshop on Deep Learning in Medical Image Analysis*, volume 10553, pages 240–248, 2017.

[150] N. N. Sultana and N. B. Puhan. Recent deep learning methods for melanoma detection: A review. In *International Conference on Mathematics and Computing*, volume 834, pages 118–132. Springer Singapore, 2018.

[151] X. Sun, J. Yang, M. Sun, and K. Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222, 2016.

[152] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, Inception-ResNet and the impact of Residual Connections on learning. In *AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[153] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[154] S. A. Taghanaki, J. Kawahara, B. Miles, and G. Hamarneh. Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification. *Computer Methods and Programs in Biomedicine*, 145:85–93, 2017.

[155] The University of Edinburgh. Dermofit image library. `https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html`, 2018. [Accessed May 22, 2019].

[156] K. Tran, M. Ayad, J. Weinberg, A. Cherng, M. Chowdhury, S. Monir, M. El Hariri, and C. Kovarik. Mobile teledermatology in the developing world: Implications of a feasibility study on 30 Egyptian patients with common skin diseases. *Journal of the American Academy of Dermatology*, 64(2):302–309, 2011.

[157] K. T. Tran, N. A. Wright, and C. J. Cockerell. Biopsy of the pigmented lesion - When and how. *Journal of the American Academy of Dermatology*, 59(5):852–871, 2008.

[158] A. Trettel, L. Eissing, and M. Augustin. Telemedicine in dermatology: Findings and experiences worldwide - a systematic literature review. *Journal of the European Academy of Dermatology and Venereology*, 32(2):215–224, 2018.

[159] P. Tschandl, G. Argenziano, M. Razmara, and J. Yap. Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. *British Journal of Dermatology*, pages 1–11, 2018.

[160] P. Tschandl, H. Kittler, and G. Argenziano. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. *British Journal of Dermatology*, 177(3):867–869, 2017.

[161] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatology*, 155(1):58–65, 2019.

[162] P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:1–9, 2018.

[163] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[164] University of Waterloo. Skin cancer detection. `https://uwaterloo.ca/vision-image-processing-lab/research-demos/skin-cancer-detection`. [Accessed May 22, 2019].

[165] T. Wadhawan, N. Situ, H. Rui, K. Lancaster, X. Yuan, and G. Zouridakis. Implementation of the 7-point checklist for melanoma detection on smart handheld devices. In *IEEE Engineering in Medicine and Biology Society*, pages 3180–3183, 2011.

[166] S. Wahie and C. M. Lawrence. Wound complications following diagnostic skin biopsies in dermatology inpatients. *Archives of Dermatology*, 143(10):1267–1271, 2007.

[167] B. N. Walker, J. M. Rehg, A. Kalra, R. M. Winters, P. Drews, J. Dascalu, E. O. David, and A. Dascalu. Dermoscopy diagnosis of cancerous lesions utilizing dual deep learning algorithms via visual and audio (sonification) outputs: Laboratory and prospective observational studies. *EBioMedicine*, 40:176–183, 2019.

[168] C. Wassef and B. K. Rao. Uses of non-invasive imaging in the diagnosis of skin cancer: An overview of the currently available modalities. *International Journal of Dermatology*, 52(12):1481–1489, 2013.

[169] J. Weingast, C. Scheibböck, E. M. T. Wurm, E. Ranharter, S. Porkert, S. Dreiseitl, C. Posch, and M. Binder. A prospective study of mobile phones for dermatology in a clinical setting. *Journal of Telemedicine and Telecare*, 19(4):213–218, 2013.

[170] World Health Organization. INTERSUN: the global UV project. `http://who.int/uv/publications/en/Intersunguide.pdf`, 2003. [Accessed May 22, 2019].

[171] World Health Organization. ICD-10. `https://icd.who.int/browse10/2016/en`, 2016. [Accessed May 22, 2019].

[172] Y. Yan, J. Kawahara, and G. Hamarneh. Melanoma recognition via visual attention. In A. Chung, J. Gee, P. Yushkevich, and S. Bao, editors, *Information Processing in Medical Imaging*, volume 11492 LNCS, pages 793–804. Springer, 2019.

[173] J. Yang and X. Sun. Recognition of clinical skin disease images. `http://cv.nankai.edu.cn/projects/sd-198/`. [Accessed May 22, 2019].

[174] J. Yang, X. Sun, J. Liang, and P. L. Rosin. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1258–1266, 2018.

[175] T. Yoshida, M. E. Celebi, G. Schaefer, and H. Iyatomi. Simple and effective preprocessing for automated melanoma discrimination based on cytological findings. In *IEEE International Conference on Big Data*, pages 3439–3442, 2016.

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

[177] Y. Yuan and Y. C. Lo. Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. *IEEE Journal of Biomedical and Health Informatics*, 23(2):519–526, 2019.

[178] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition Learning*, pages 2921–2929, 2016.

[179] M. H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577, 1993.