

University of Modena and Reggio Emilia

XXXIV cycle of the international doctorate school in
Information and Communication Technologies

Doctor of Philosophy dissertation in
Computer Engineering and Science

Deep Learning in Dermoscopic Image Analysis

Federico Pollastri

Advisor: Prof. Costantino Grana

PhD School Director: Prof. Sonia Bergamaschi

Modena, 2022

Abstract

Countless different imaging acquisition techniques are employed by medical practitioners as a tool to perform diagnosis, ranging from microscopy to Magnetic Resonance Imaging (MRI). This common practice produced a great opportunity for computer vision algorithms to find ways to perform automated analysis on medical images. In particular, after the groundbreaking success of AlexNet in 2012, deep learning has become a vital element in medical imaging research. Above all, Convolutional Neural Networks (CNNs) have been successfully adopted to perform a great variety of tasks such as image segmentation, classification, detection, and generation. This thesis is a collection of deep learning applications for dermoscopic images analysis. Dermoscopy is a form of in-vivo skin surface microscopy performed using high quality magnifying lenses and a powerful light source to mitigate the surface reflection of the skin, to enhance the visibility of the pigmentation of the lesion. However, to fully make use of this non-invasive imaging approach, a thorough image analysis must be performed by expert clinicians, and therefore many efforts have been given in recent years towards the creation of tools to assist physicians in the analysis of dermoscopic images. We start by approaching lesion segmentation, by means of a novel data augmentation technique and a diverse ensemble strategy. The final goal of dermoscopic images analysis is skin lesion classification, for which we develop an approach that achieved the third best result in the 2019 ISIC global challenge. Moreover, we address one of the main drawbacks of deep learning algorithms, their low interpretability, by using content-based image retrieval to assist the diagnosis process of both expert and novice practitioners and, finally, by trying to determine which characteristics are taken into account by autonomous classification algorithms.

Abstract (Italian)

In medicina, innumerevoli tecniche di acquisizione di immagini impiegate come strumento per eseguire diagnosi, dalla microscopia alla risonanza magnetica (MRI). Questa comune pratica ha prodotto una grande opportunità per gli algoritmi di visione artificiale di trovare modi per eseguire analisi automatizzate di immagini mediche. In particolar modo, dopo il rivoluzionario successo di AlexNet nel 2012, il deep learning è diventato un elemento vitale nella ricerca sulle immagini mediche. Più precisamente, le reti neurali convoluzionali (CNN) sono state adottate con successo per affrontare una grande varietà di problemi come la classificazione e la generazione di immagini, o l'identificazione e la segmentazione di oggetti all'interno di esse. Questa tesi è una raccolta di applicazioni di deep learning su immagini dermoscopiche. La dermoscopia è una forma di microscopia in vivo della superficie cutanea, eseguita utilizzando lenti di ingrandimento di alta qualità e una potente fonte di luce, che mitiga il riflesso superficiale della pelle con lo scopo di migliorare la visibilità della pigmentazione della lesione. Tuttavia, per utilizzare a pieno questo approccio di imaging non invasivo, è necessaria un'analisi approfondita dell'immagine da parte di dermatologi esperti, e la comunità scientifica ha quindi impiegato grandi risorse per la creazione di strumenti automatici che possano assistere l'analisi di immagini dermoscopiche. Affrontiamo per prima la segmentazione delle lesioni cutanee, mediante una nuova tecnica di data augmentation e una strategia di insieme diversificata. L'obiettivo finale dell'analisi delle immagini dermoscopiche è la classificazione delle lesioni, per la quale sviluppiamo un approccio che ha ottenuto il terzo miglior risultato nella sfida globale ISIC 2019. Inoltre, affrontiamo uno dei principali svantaggi degli algoritmi di deep learning, la loro bassa interpretabilità,

utilizzando un recupero di immagini basato sul contenuto per assistere il processo di diagnosi sia di dermatologi esperti che di principianti. Infine, cerchiamo di determinare quali caratteristiche siano prese in considerazione dagli algoritmi di classificazione automatica.

Review committee composed of:

Giuseppe Serra

University of Udine

Cosimo Distanto

National Research Council (CNR)

University of Salento

The difference between science and philosophy is that experiments can show that extremely plausible ideas are just wrong and extremely implausible ones, like learning a entire complicated system by end-to-end gradient decent, are just right.

Geoffrey Hinton

Contents

Contents	xi
1 Introduction	1
2 Literature Survey	3
2.1 Augmenting data with GANs	3
2.2 Image Segmentation	4
2.3 Skin Lesion Classification	5
2.4 Content-Based Dermoscopic Image Retrieval	6
2.5 Neural Networks Interpretability	7
3 Augmenting data with GANs to segment melanoma skin lesions	9
3.1 Introduction	9
3.2 Data Augmentation with GANs	11
3.2.1 DCGAN	11
3.2.2 LAPGAN	12
3.2.3 Samples Quality Discussion	13
3.3 Dermoscopic Image Segmentation	15
3.3.1 Baseline Architecture	15
3.3.2 Hyperparameters Analysis	15
3.4 Experimental Results	17
3.4.1 About Flexibility	19
3.4.2 Overall Accuracy	20
3.4.3 About Scalability	22
3.5 Conclusion	25

4	Skin Lesion Segmentation Ensemble with Diverse Training Strategies	27
4.1	Introduction	27
4.2	Learning to Extract Features	28
4.2.1	ISIC Task 3	28
4.2.2	Autoencoder	29
4.2.3	GAN	30
4.3	Learning to Segment Skin Lesions	31
4.3.1	Detection-Segmentation Pipeline	32
4.3.2	The Ensemble Method	33
4.4	Experimental Results	34
4.5	Conclusion	36
5	A Deep Analysis on High Resolution Dermoscopic Images Classification	39
5.1	Introduction	39
5.2	Dermoscopic Images	41
5.2.1	Preprocessing	42
5.3	High Accuracy Through Growth	48
5.3.1	Probabilistic Model	48
5.3.2	Averaging Calibrated Probabilities	53
5.4	Experimental results	54
5.5	Conclusion	60
6	Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval	61
6.1	Introduction	61
6.2	Image Retrieval	63
6.2.1	Classification	63
6.2.2	Embedding End-to-End	64
6.2.3	Classification & Embedding	65
6.2.4	Classification & Embedding End-to-End	65
6.3	Experimental Results	66
6.3.1	Quantitative	66
6.3.2	Qualitative	71
6.4	Conclusion	72

7	Where Do Neural Networks Look?	73
7.1	Introduction	73
7.2	Dermoscopic Images	75
7.2.1	Private Dataset	76
7.3	Investigating ABCDE features	77
7.3.1	Tampering with Asymmetry	78
7.3.2	Concealing Borders	79
7.3.3	Removing Color	79
7.3.4	Altering Diameter	80
7.3.5	About Evolving	80
7.4	Experiments	81
7.5	Discussion	86
7.5.1	A Crucial Characteristic in Dermoscopic Datasets	87
7.5.2	About AUC	88
7.6	Conclusion	93
8	Conclusion	95
	Appendix	97
A	Additional Experimental Results	97
B	List of Publications	103
C	Activities Carried Out During the PhD	106
	Bibliography	111

Acknowledgments

I would like, first of all, to acknowledge my advisor Prof. Costantino Grana for all of his efforts to guide and assist me during my PhD. Thank you for inspiring and motivating me to always pursue high-quality scientific results, having you as an advisor has made all the difference in the world.

I would like to express my gratitude to the whole AImageLab research group, to the Department of Engineering “Enzo Ferrari”, and to the “Information and Communication Technologies” doctoral school, for giving me the opportunity to be part of this graduate program and support me throughout. A special thank you to my lab colleagues Federico, Michele, Stefano, Laura, and Marco for being exceptional co-workers and most importantly great friends, even more so in the months we were forced apart by global events.

None of my accomplishments would have been possible without the assistance of my family. Thank you for giving me the chance to choose my own course, for teaching me the relevance of hard work, and for your endless support. Thank you to my girlfriend, for all of her love and for the constant encouragement. Thank you to all of my friends, who have been staying by my side for a very long time.

Chapter 1

Introduction

Malignant melanoma is the most dangerous type of skin cancer, with a substantial death rate. It can be cured with prompt excision if detected in the early stage, making fast diagnosis extremely important [1]. However, early detection is very hard to obtain, and even in specialized centers the accuracy of the clinical diagnosis for melanoma, achieved with the unaided eye, is slightly better than 60% [2]. To perform a fast diagnosis, many dermatologists rely on dermoscopy, which is a form of in-vivo skin surface microscopy performed using high quality magnifying lenses and a powerful light source to mitigate the surface reflection of the skin, in order to enhance the visibility of the pigmentation of the lesion. This imaging technique has increased the diagnosis accuracy, sensitivity, and specificity with respect to the naked eye examination, mitigating the need of surgical intervention for the unnecessary removal of benign lesions. However, to diagnose skin cancer through this kind of non-invasive imaging approaches, a thorough image analysis must be performed by expert clinicians. This is why many efforts have been given in recent years towards the creation of tools to assist physicians in the analysis of dermoscopic images. Deep learning in particular, due to its outstanding results in many areas such as speech recognition [3], image understanding [4] and image classification [5, 6], has become the main option for analyzing medical images.

Unfortunately, in order to be trained, Convolutional Neural Networks (CNNs) require huge amounts of data, which are often hard to obtain and particularly expensive to annotate, especially in medical fields. This

is why the International Skin Imaging Collaboration (ISIC) has begun to aggregate a large-scale, publicly accessible dataset of dermoscopic images. This dataset enabled the development of the first public benchmark challenge on dermoscopic image analysis in 2016. The goal of the challenge was to provide a fixed dataset snapshot to support development of automated melanoma diagnosis algorithms across three tasks of lesion analysis: segmentation, dermoscopic feature detection, and classification. The hunger for large amounts of annotated medical data can be mitigated by pre-training neural networks with an already existing collection of natural images, like ImageNet [7]. Indeed, the first convolutional layers of every CNN learn to recognize simple elements like lines and colors, making them useful across different tasks and datasets [8]. Learning features from wider datasets and re-using them in different tasks can be seen as a form of transfer learning. The choice of the dataset employed to learn low-level features can introduce biases towards certain features and characteristics, and this should be taken into account especially when dealing with medical imaging. For example, CNNs trained using ImageNet are strongly biased in recognizing textures rather than shapes [9].

With this thesis, we first focus on the segmentation task, a fundamental step in the automated melanoma detection process, defined as the recognition of the set of pixels that constitute the skin lesion within the image. We approach it by first developing a novel pre-training strategy in Chapter 3 and then merging it with additional state-of-the-art techniques in a segmentation ensemble in Chapter 4. In Chapter 5 we dive into the classification task, performing a comprehensive analysis on the effectiveness of state-of-the-art deep learning techniques when applied to skin lesion classification. In order to achieve this goal, we consider several CNN architectures and analyze how their performance is affected by image resolution, size of the network, data augmentation process, amount of available data, and model calibration. In order to build a tool to boost the classification accuracy of human dermatologists, in Chapter 6 we design and develop a novel skin image retrieval system, which exploits features extracted by Convolutional Neural Networks to gather similar images from a publicly available dataset, in order to assist the diagnosis process of both expert and novice practitioners. Finally, in Chapter 7, we explore the interpretability of CNNs in dermoscopic image analysis, and try to determine which characteristics are taken into account by autonomous classification algorithms.

Chapter 2

Literature Survey

2.1 Augmenting data with GANs

Deep learning algorithms, Convolutional Neural Networks (CNNs) in particular, have become the methodology of choice for analyzing medical images. One of the main reasons is most certainly their ability to extract features on their own [10], as opposed to many previous approaches based on hand-crafted features [11, 12, 13]. In classical Convolutional Neural Networks, convolutions and non-linearities are interleaved with pooling operations. The purpose of these subsampling units is to enlarge the convolved receptive field and to decrease the feature maps size. These two effects are respectively important to obtain information from wider areas for the final prediction, and to augment the number of feature maps without the memory requirements becoming excessive. Unfortunately, resolution is a key element in segmentation applications, where every pixel is expected to be assigned to a specific class.

An important operation which should precede the network training is data augmentation. This process consists in generating new data items by applying very simple transformations to existing training samples [14]. Data augmentation is very often exploited in deep learning to artificially enlarge a dataset without needing new data and, most importantly, without compromising the consistency of the training set. In skin lesion segmentation, a huge variety of data augmentation steps can be applied to every single image, including random rotation, flipping, shifting and scaling [15].

Instead of relying on manually defined transformations, it would be nice to learn how to augment data in the same deep-learning framework. Generative Adversarial Networks [16] are a model made up by two different networks: a generator and a discriminator. The task of the discriminator is learning to distinguish real samples from generated ones, while the generator tries to generate good enough samples to fool the discriminator. The notable effect of this architecture is that the two networks improve together, learning at the same time while competing. Neff *et al.* [17], for instance, employed GANs to generate annotated images in order to simply mix them with the original data, overlooking the influence of the generated samples' quality. Moreover, the work by [18] explores the possibility of exploiting GANs to generate skin lesion images. That paper focuses on the generation of skin lesion images, tackling the heavy class imbalance that afflicts the classification task, without including any segmentation mask.

2.2 Image Segmentation

In the last decades, a lot of effort has been devoted to solving the problem of automated skin lesion segmentation, promoting the development of different approaches: histogram thresholding [19, 20], edge-based [21, 22], region-merging [23, 24], and supervised learning [25]. Nevertheless, CNNs are now one of the most powerful tools in machine learning and computer vision. They are able to learn features directly from input image data, with no need of hand-crafted features. In a typical classification task a convolutional network produces a single label as output. However, in other visual tasks such as segmentation [26, 27], a pixel-wise output information is required, having a critical point in the mapping of low resolution features into input resolution.

In 2015, Ronneberger *et al.* proposed the so called U-Net [28], which consists of a contracting encoder network that extracts high resolution features from the input image and followed by an expanding decoder model to produce a full-resolution segmentation. This model employs skip connections between the downsampling and upsampling path, applying a concatenation operator in order to provide coarse contextual information when upsampling.

In DeepLabv1 [29], atrous convolutions were introduced. Instead of repeatedly reducing the dimension of feature maps, each convolutional fil-

ter of the network is upsampled in size and filled with zeroes in between filter values. This allows to avoid the use of pooling operations, which are responsible for the spatial resolution reduction, and thus cause a lower output segmentation quality.

DeepLabv3+ [30] is a state-of-the-art CNN for semantic segmentation. It employs atrous convolutions, residual blocks [5], and a decoder module to refine the segmentation results.

In 2017, Mask R-CNN [31] became the main baseline for instance segmentation, which differs from semantic segmentation in the requirement of identifying each distinct object individually. This approach combines object detection, which aims to classify and localize individual objects using a bounding box, and semantic segmentation, which has the goal of classifying each pixel given a fixed set of classes. This strategy extends Faster R-CNN [32], adding a Fully Convolutional Network (FCN) to predict segmentation masks for each Region of Interest (RoI) in a pixel-to-pixel fashion.

Newer algorithms have been proposed building on active contours [33], self-training [34], densely connected neural architecture search [35], and hierarchical multi-scale attention [36]. However, despite the minor accuracy improvements achieved after their proposal, both DeepLabv3+ and Mask R-CNN remain state-of-the-art architecture for semantic and instance segmentation respectively [37].

2.3 Skin Lesion Classification

CNNs have become the dominant machine learning approach, and the *scaling up* strategy has been widely used to achieve better accuracy results. As an example, ResNet [5] can be scaled up from ResNet-18 to ResNet-200 just by adding more layers. However, this technique leads to the notorious problem of vanishing/exploding gradients [38], which hampers the convergence of the architectures. This problem has been managed with different approaches such as intermediate normalization layers [39] or normalized initialization [40, 41], resulting in great accuracy improvements over the years. In 2016, Xie *et al.* proposed ResNeXt [42], which introduces the concept of *cardinality*, slightly changing the residual block structure proposed with ResNet. In the same year, DenseNet [43] proposed a new architecture which increases the number of connections of each layer, al-

leviating the vanishing-gradient problem. SEResNeXt [44], introduced in 2017, provides significant performance improvements w.r.t. existing state-of-the-art CNNs, by means of “Squeeze-and-Excitation” (SE) blocks, that adaptively recalibrate channel-wise feature responses. Finally, in 2019 EfficientNet [6] provided a new scaling up method that uniformly increases the dimensions of depth, width, and resolution, achieving state-of-the-art accuracy on ImageNet [7].

Transfer Learning has become a *de-facto* method to enhance the training process of deep learning models. It is a vital tool to be exploited when the data in the target domain is not abundant. Natural image datasets such as ImageNet are usually chosen to pre-train neural networks, due to the vast amount of labeled data. However, the performance can worsen when the source and target domains present several differences [45, 46]. Fortunately, it is frequently possible to take advantage of these pre-trained features and increase the network accuracy thanks to the transferability of simpler filters, such as those that address color, size, or edges [47].

Furthermore, the performance of a CNN can be boosted through an extensive investigation of the hyperparameterization and using ensembles of several models [48, 49]. The latter approach is stated to produce the best accuracy results [50], despite a heavy computational cost in terms of resources, training time, and inference time. However, in the skin lesion analysis domain, reliable results must be preferred to low inference time.

2.4 Content-Based Dermoscopic Image Retrieval

In the work by Ballerini *et al.* [51], the authors developed a content-based image retrieval system for a dataset of 533 images, divided into five classes of lesions, including two non-melanoma cancer types. Their system relies on visual features, such as color and composite texture, evolved using genetic algorithms. The similarity matching function is obtained by composing Bhattacharyya distance [52] for color covariance-based features, and Euclidean distance for texture features. In [53], Baldi *et al.* proposed another method based on low-level representative features. They find visually similar images to a query by means of a hierarchical multi-scale computation of the Bhattacharyya distance of all the database images. Another CBIR system for skin lesions was developed by Jiji and Raj [54], who ex-

exploit features such as color, shape and texture, in order to find the most visually similar images to a query in a dataset composed of 20 different diseases. Their categories, however, do not include melanocytic nevi, and melanomas are not distinguished from other kinds of skin cancer. In 2016, Rahman *et al.* [55] proposed a decision support system based on a fusion between classification and retrieval. Their model considered high-level features such as Non-Subsampled Contourlet Transform (NSCT) and HOG based on Hessian matrix¹. The model has been validated on the ISIC 2016 dataset, containing two skin lesion categories, melanoma and benign. Belattar *et al.* [56] used a kernelized SVM classification algorithm with an active learning technique and a histogram intersection matching in a retrieval system with relevance feedback for melanoma diagnosis. Their experiments were conducted on a dataset of melanomas and benign lesions. Finally, Pu *et al.* [57] applied the retrieval method proposed in [58] to the skin lesion domain. Their approach consists of training a deep network that learns to map images to hash codes, in a way that preserves similar semantics. Their network minimizes a classification loss function, with additional regulation terms to achieve desirable hash code properties. Then, similar images to a query are ranked by means of the Hamming distance, and clustered with Affinity Propagation (AP). This last proposal improves the measured precision with respect to previous methods based on hand-crafted features, and therefore represents a reference for comparison.

2.5 Neural Networks Interpretability

Although deep neural networks show great performance, their lack of decomposability into comprehensible components makes them hard to interpret. The low outcome explainability inevitably generates the risk of training neural networks to classify images based on trivial artifact, instead of meaningful features, as recently proved by several COVID-19-related publications [59, 60]. Maguolo *et al.* analyzed diverse COVID-19 datasets also hiding sensitive information with large black squares concluding that “these protocols might be biased and learn to predict features that depend more on the source dataset than they do on the relevant medical information”. Robert *et al.*, on the other hand, carried out an extensive analysis of the risk of biases in state-of-the-art methods for COVID-19 identification

¹Eig(Hess)-HOG

in CT and CXR (standard-of-care chest radiographs) scans. The authors claim that “none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases”.

Similar observations have been made about dermoscopic image analysis by Bissoto *et al.* [61], who recently published a study to show the effects of performing skin lesion classification while occluding the actual skin lesion with large black bounding boxes. The melanoma/non-melanoma classification performance reached an AUC (Area Under the ROC Curve) of 77.4%, which is quite inferior compared to state-of-the-art methods, but higher than what expert dermatologists can do [62]. On the other hand, separate studies [63, 64] showed that CNNs are able to learn filters sensitive to both important features, such as lesion borders and darker regions inside the lesion, and meaningless characteristics, like surrounding skin and image artifacts.

Autonomous systems in medical applications aim to act as support tools for clinicians, and therefore must be trustworthy and highly interpretable. Outcome explainability of neural networks can be increased thanks to several visualization strategies such as *CAMs* (Class Activation Mapping), which have been proposed for the identification of image regions which most contribute to the final prediction [65, 66].

Chapter 3

Augmenting data with GANs to segment melanoma skin lesions

3.1 Introduction

The skin lesion segmentation task can be especially complicated due to the great variety of characteristics shown in skin lesions and skin itself among different people (Fig. 3.1). The definition of skin lesion borders is also very subjective. Moreover, manual image segmentation is a very time consuming job that requires the work of a competent specialist. Segmentation CNNs are able to extract features through a contracting path and exploit them to generate a segmentation mask across the expanding path; the size of feature maps decreases progressively in the former, whereas in the latter it increases back to the input resolution thanks to up-sampling operators and fractionally-strided convolutions, thus producing an encoder-decoder architecture. State-of-the-art algorithms for lesion segmentation are based on supervised machine learning techniques, it is then crucial to find a cheaper way to obtain segmented images useful to train the models.

Therefore, the focus of this chapter is the data augmentation process of a Convolutional-Deconvolutional Neural Network (CDNN), designed to automatically map dermoscopic images into lesion segmentation masks. In

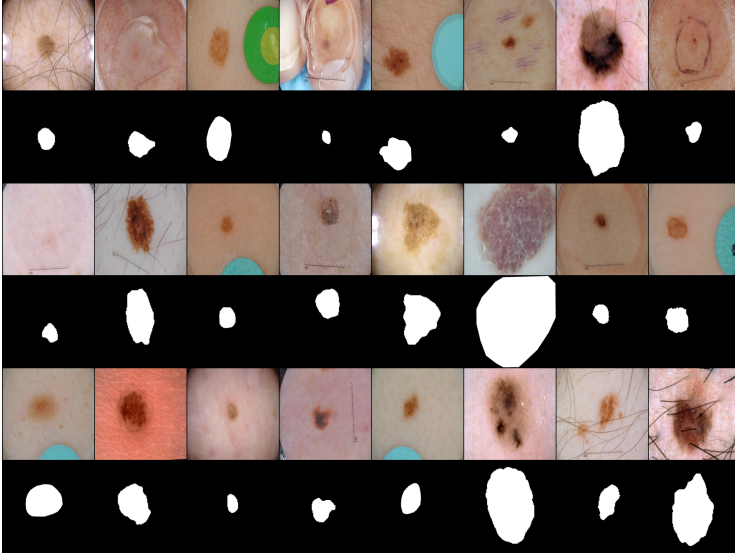


Figure 3.1: Samples from the ISIC dataset: dermoscopic skin images coupled with their ground truth segmentation masks.

particular, we propose a framework to generate both skin lesion images and their segmentation masks by means of a Generative Adversarial Network (GAN). Differently from the usual approach, in which GANs are used to generate unlabeled images, tricky to include in the training process, the ability to reproduce image-mask couples makes the data augmentation process extremely straightforward [67]. Indeed, data generated in this way can be used to perform an initialization of deep networks, always crucial in deep learning [68]. After some epochs, original samples are used to fine-tune the network parameters.

Experimental results show that adding GAN-generated data in the training process effectively improves the segmentation accuracy of state-of-the-art CDNNs.

The rest of the chapter is organized as follows: Section 3.2 summarizes two different GANs employed to generate synthetic images, whereas Section 3.3 focuses on the automated segmentation model that will benefit from said generated data. The training process and experimental results of

our CDNNs are presented in Section 3.4. In this Section, the effectiveness of the GAN-generated data on a shrunk dataset are also explored. Finally, in Section 3.5 conclusions are drawn.

3.2 Data Augmentation with GANs

The segmentation task could greatly benefit from a bigger and wider dataset, this is why we approach the problem focusing on the data augmentation process. GANs are often used to create unlabeled examples [69, 70, 71], which cannot be directly employed for the training of a supervised algorithm. We improve the role of the GAN in the training process by implementing an architecture which generates both the image and its segmentation mask, making it extremely easy to exploit new synthetic images as additional training data. In the proposed GANs, both the input and the output are 4 channel images: the first three channels are the R, G and B components and the fourth one is the binary segmentation mask. This approach provides a tool for generating new skin lesion images coupled with a coherently learned segmentation mask. In order to train our GANs, we choose to remove the 118 non-dermoscopic images from the 2017 ISIC challenge training set, given their very different nature. This approach sets the final training dataset size to 1882 images.

3.2.1 DCGAN

We modify the DCGAN proposed in [72] in order to use the described 4D input/output data in both the generator and the discriminator.

The two networks exploit four convolutional/deconvolutional layers, with a kernel size of 5×5 , and a stride of two. Each layer, but the last two, is followed by a Leaky Rectified Linear Unit, LReLU [73], and batch normalization [39]. A sigmoid activation function is added after the fully-connected layer at the end of the discriminator to perform the binary classification (real or fake). A hyperbolic tangent serves as the activation function of the generator. After the training process is completed, our generator is able to employ 100-dimensional random vectors to create 192×256 RGB synthetic skin lesion images and their 192×256 binary segmentation masks (Fig. 3.2a and Fig. 3.2b).

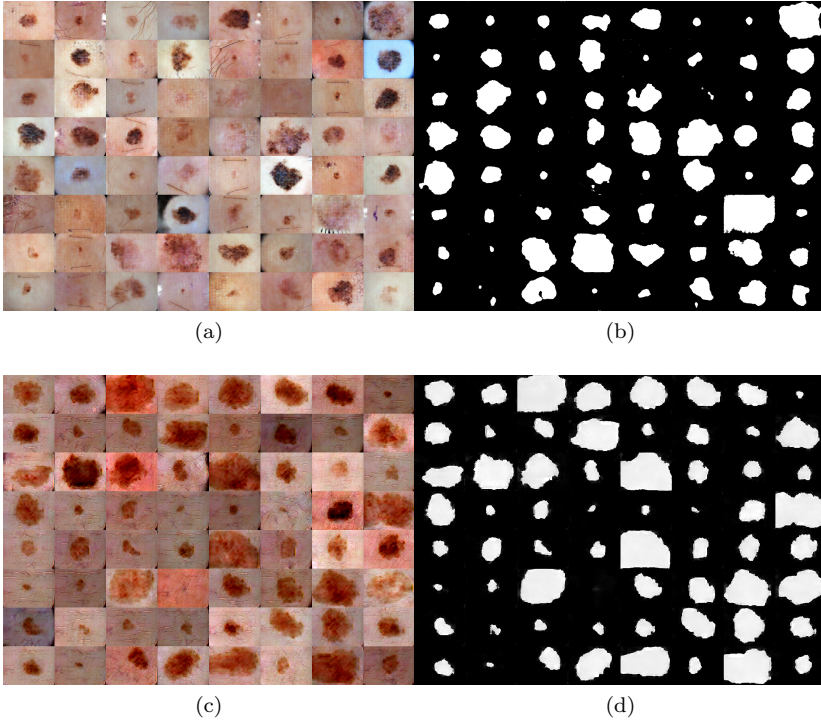


Figure 3.2: DCGAN-generated skin lesion samples (a) and their segmentation masks (b). LAPGAN-generated skin lesion samples (c) and their segmentation masks (d).

3.2.2 LAPGAN

We implement a 5 levels Laplacian Generative Adversarial Network (LAPGAN) by designing and training 5 different, independent GANs [74]. The network representing the first level of the laplacian pyramid generates a 6×8 pixels dermoscopic image from a 100-dimensional random vector. In this case, both the generator and the discriminator are formed by two fully connected layers. The following pyramid levels have a different task: they learn how to improve the resolution of upsampled and blurred images,

and each one focuses on images with a different fixed size. The generators of the four last GANs are three layers CNN, whereas discriminators are composed by two convolutional layers and a fully connected one.

LReLU activation functions serve after each intermediate layer, while the output of the discriminators is affected by a sigmoid function. No activation function is added to the last layer of the generators.

Starting from the 6×8 pixels data, each image is upsampled and fed into the next pyramid level, following the sampling process illustrated in the appendix. An extra iteration with the last GAN is implemented to obtain 192×256 images and their segmentation masks (Fig. 3.2c and Fig. 3.2d).

3.2.3 Samples Quality Discussion

Both the DCGAN and the LAPGAN are trained to generate 192×256 images and segmentation masks.

The DCGAN benefits from its end-to-end training process, learning to produce visually pleasing skin lesion samples, with good diversity. Most images present realistic details, like the well delivered presence of hair, black corner representing real camera characteristics or pen marks. The images, however, suffer from a heavy checkerboard effect caused by the fractional strided convolution layers implemented in the framework (Fig. 3.3a).

The LAPGAN is able to get rid of the checkerboard effect as a result of the absence of previously described layers. On the other hand, this model introduces noisy lesion borders and artifacts such as unnatural square green patches and low quality hair (Fig. 3.3b), and images generated by the LAPGAN look much similar one to another. Moreover, the particular nature of a LAPGAN makes the training process extremely hard: the five different required networks need constant supervision and hyperparameters adjustment, making good results very painful to obtain, especially when compared to the DCGAN.

Overall, both models produce segmentation masks that look very coherent with the respective generated images. We thus proceed to evaluate the impact of the two frameworks on the accuracy of lesion segmentation architectures, when they are included in the training process.

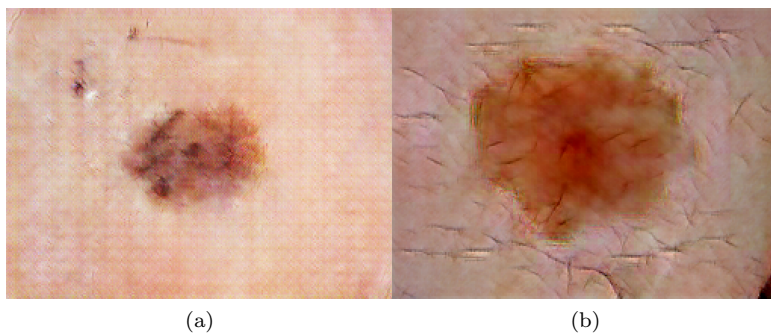


Figure 3.3: (a) detail image of the DCGAN output, and (b) detail image of the LAPGAN output.

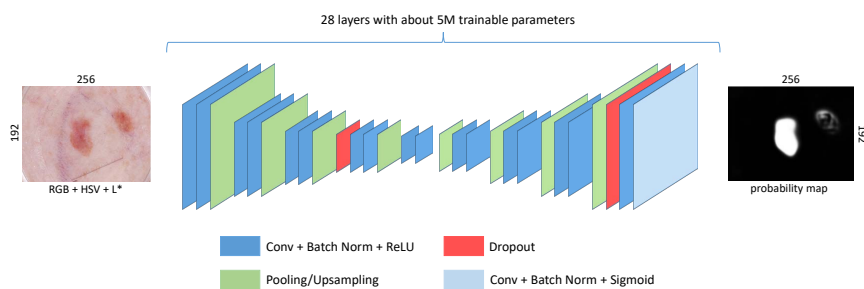


Figure 3.4: Baseline CDNN model.

3.3 Dermoscopic Image Segmentation

3.3.1 Baseline Architecture

In order to avoid trivial comparisons, we select and reimplement the baseline CDNN from the architecture that obtained the highest score in 2017 ISIC challenge [15]. This network maps the input dermoscopic image to a posterior probability map. The network contains 28 layers with about 5M trainable parameters. The stride is fixed at 1 and Rectified Linear Units are used as activation functions for each convolutional/deconvolutional layer but the output, which implements a sigmoid.

As in the original proposal, every image is resized to 192×256 and the original RGB channels are augmented with both the HSV and L^* channels. In accordance with the original paper, data augmentation is obtained by randomly flipping, rotating, shifting, scaling and changing color contrasts in original images and ground truths.

The loss function designed by [15] is:

$$L = 1 - \frac{\sum_{i,j} t_{ij} p_{ij}}{\sum_{i,j} t_{ij}^2 + \sum_{i,j} p_{ij}^2 - \sum_{i,j} t_{ij} p_{ij}}, \quad (3.1)$$

where t_{ij} is the target value of the pixel at coordinates (i, j) , and p_{ij} is the real output. Note that t_{ij} is either 0 or 1, while p_{ij} is a real number in range $[0, 1]$.

Ensemble methods require different techniques to be combined and used together, so we choose to introduce variations in the original CDNN (Fig. 3.4) by studying the main hyperparameters of the model. The final purpose is to obtain many networks different from one another, and then combine their output through a simple ensemble method. These networks, named CDNN_0 through CDNN_6 , are summarized in Table 3.1 and described in the following section.

3.3.2 Hyperparameters Analysis

The baseline architecture (CDNN_0) is mainly affected by three hyperparameters: input image size, color channels and loss function. We describe

Table 3.1: Hyperparameters of the Neural Networks analyzed.

NN	Images Dimension	Images Channels	Loss
CDNN ₀	192×256	R, G, B, H, S, V, L*	Eq. 3.1
CDNN ₁	192×256	R, G, B	Eq. 3.1
CDNN ₂	192×256	R, G, B, H, S, V, L*, a*, b*	Eq. 3.1
CDNN ₃	96×128	R, G, B, H, S, V, L*	Eq. 3.1
CDNN ₄	384×512	R, G, B, H, S, V, L*	Eq. 3.1
CDNN ₅	192×256	R, G, B, H, S, V, L*	Eq. 3.3
CDNN ₆	192×256	R, G, B, H, S, V, L*	Eq. 3.4

a set of possible variations over these parameters (Table 3.1), in order to stress how they influence the final performance results.

Image Channels. In the baseline model both HSV and L* channels were added. These channels are obtained by means of non-linear transformations. To address the question whether the model is able to independently learn these transformations autonomously, we train two networks: one with just the three RGB channels (CDNN₁) and one with every channel from RGB, HSV and CIELAB spaces (CDNN₂).

Resizing Dimensions. We train two additional networks with input images resized to 96×128 and 384×512 (CDNN₃ and CDNN₄), which are respectively half and double the original size along each axis. Since we do not change the stride and the size of convolutional filters, the scaling factor between layers remains the same. This provides the two networks with a different encoded representation size.

Loss Function. It is interesting to point out the common misconception displayed in the work by [15]. The distance measure in Eq. 3.1 is said to be based on the Jaccard distance. In fact, Eq. 3.1 is the Tanimoto distance, which is a proper distance when both vectors have only positive elements, and it is equal to the Jaccard distance only with binary vectors [75]. In our case, only the target is binary, but the prediction is a real value between 0 and 1. The correct (generalized) Jaccard distance on real positive vectors is defined as

$$d_J = 1 - \frac{\sum_{i,j} \min(t_{ij}, p_{ij})}{\sum_{i,j} \max(t_{ij}, p_{ij})}. \quad (3.2)$$

Since t_{ij} is still binary, Eq. 3.2 can also be computed as

$$d_J = 1 - \frac{\sum_{i,j} t_{ij} p_{ij}}{\sum_{i,j} t_{ij} + \sum_{i,j} p_{ij} - \sum_{i,j} t_{ij} p_{ij}}, \quad (3.3)$$

which may be the reason for the common confusion (note the missing squares with respect to Eq. 3.1).

We thus train two more variations of the model. The first one using the proper Jaccard distance and the other one using the mean squared error function (CDNN₅ and CDNN₆):

$$MSE = \frac{1}{n} \sum_{i,j} (t_{ij} - p_{ij})^2 \quad (3.4)$$

where n is the total number of pixels.

3.4 Experimental Results

In order to employ GAN-generated images discussed in Section 3.2, we design a two-step training process. In the first phase, each CDNN is trained with synthetic data for a total of 400 epochs¹. Instead of generating a fixed number of samples, GANs are required to provide new couples image-mask for each training batch. In the second step, the network is fed with real data, for a total of 100 epochs (Fig. 3.5).

This learning process proved itself to be the best way to exploit such kind of GAN-generated data: further experiments show that mixing real and synthetic images in a single training process reduce the accuracy of 0.020 on average.

In order to prove the effectiveness of the GANs on the final accuracy, we also train a version of each CDNN using only real data, for a total of 400 epochs.

Following the approach introduced in Section 3.2, we train each CDNN with the 1882 dermoscopic images of the 2017 ISIC challenge training set, choosing to remove the 118 non-dermoscopic ones. The batch size is set to 16 and the learning rate is fixed as 0.0003 at the beginning of each

¹An epoch is the number of batches needed to feed the network with 1882 images.

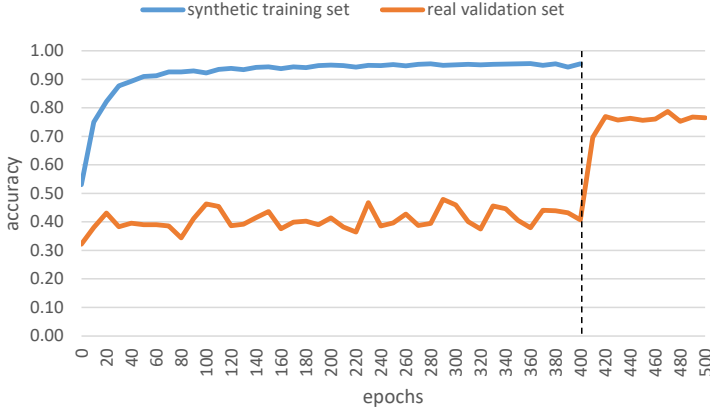


Figure 3.5: Training process of the CDNN_0 exploiting both synthetic (before the dotted line) and real data (after the dotted line). The blue line is the accuracy obtained on the synthetic training set, whereas the orange line identifies the accuracy obtained on the real validation set.

training process and it is then affected by the Adam optimizer [76]. Given the importance of data augmentation, both real and synthetic images are randomly rotated, flipped, shifted, and scaled. The color contrast is also changed.

Experimental results are summed up in Table 3.2. The first column identifies each network with a reference to Table 3.1, whereas every other one reports the Jaccard Index² obtained on the 2017 ISIC challenge public test set.

The column *Original Data* shows the accuracy of the networks trained with only real dermoscopic images. Most of the variations explored with the hyperparameters analysis obtain results close to our baseline network (CDNN_0). The only case where an actual difference is noticeable is the CDNN_4 network, the one with larger images, which obtains a Jaccard index of 0.717. Because of its low performance, it is removed from further analysis. The last three columns show the results obtained after including the two GANs in the training process. The heavy checkerboard effect, noticeable in samples generated using a DCGAN, has a negative impact

²The official accuracy measure of the ISIC challenge [77].

Table 3.2: Comparison of the experimental results on $CDNN_0$ through $CDNN_6$, when trained exploiting different augmentation strategies. Bold values represent the best performing approaches for the networks included in the ensemble method described in Section 3.4.2.

NN	Original Data	DCGAN Augmented	LAPGAN Augmented	Both GANs Augmented
$CDNN_0$	0.746	0.751	0.739	0.742
$CDNN_1$	0.759	0.753	0.764	0.752
$CDNN_2$	0.746	0.756	0.762	0.750
$CDNN_3$	0.755	0.762	0.764	0.756
$CDNN_4$	0.717	—	—	—
$CDNN_5$	0.739	0.752	0.740	0.747
$CDNN_6$	0.731	0.749	0.756	0.751

on the model accuracy.

Images generated through the LAPGAN get the best response from the architecture, despite the low visual quality and variability offered. Finally, the last column of Table 3.2 displays that alternating the two GANs during the first training phase has almost no positive outcome.

3.4.1 About Flexibility

In order to further analyze the proposed augmentation method, we choose to investigate the effect of GAN-generated samples on the training process of U-Net [28], a CDNN specifically developed for the segmentation of bio-medical images. Given the ability of this architecture to achieve good performances with only few annotated images available for training, U-Net is considered to be an excellent baseline for many medical image segmentation tasks.

The standard U-Net loss is the cross-entropy:

$$CE = - \sum_{i,j} t_{i,j} \log(p_{i,j}) \quad (3.5)$$

which, when the network is outlined to predict only two classes (background and foreground), can be also defined as Binary Cross-Entropy:

$$BCE = - \sum_{i,j} t_{i,j} \log(p_{i,j}) + (1 - t_{i,j}) \log(1 - p_{i,j}) \quad (3.6)$$

Table 3.3 summarizes the results obtained by using U-Net with two different losses: U-Net standard loss (Eq. 3.6) and our baseline loss (Eq. 3.1), validating the superiority of LAPGAN-generated samples highlighted in Section 3.4, and proving the Binary Cross-Entropy to be the best performing loss for U-Net.

Table 3.3: Comparison of the experimental results on U-Net, when trained exploiting different augmentation strategies. Bold values represent the best performing approaches.

Loss	Original Data	DCGAN Augmented	LAPGAN Augmented	Both GANs Augmented
Eq. 3.6	0.753	0.755	0.767	0.764
Eq. 3.1	0.744	0.755	0.762	0.760

3.4.2 Overall Accuracy

For each pixel, the probability of it being part of the skin lesion is obtained as the mean value across the selected CDNNs (*i.e.* all the networks presented in Section 3.3.2 but CDNN₄). The output is then binarized with a dual-threshold method. A high threshold (0.8) is followed by Connected Components Labeling [78, 79] and the biggest object center is assumed to be the tumor center. Afterwards, a lower threshold (0.4) is applied and the final segmentation mask is given by the region which contains the tumor center, as shown in Fig. 3.6. When the first high threshold does not find any object, we only apply the second one and keep its result as segmentation mask.

It is interesting to point out that, even though the accuracy of the various CDNNs is almost uniform, different networks specialize on different features, improving the segmentation accuracy on some images rather than on others (Fig. 3.6). This greatly increases the effectiveness of our ensemble method.

CDNN₃ is the only network that, even with great results obtained when trained with the original data, always benefits from the proposed data aug-

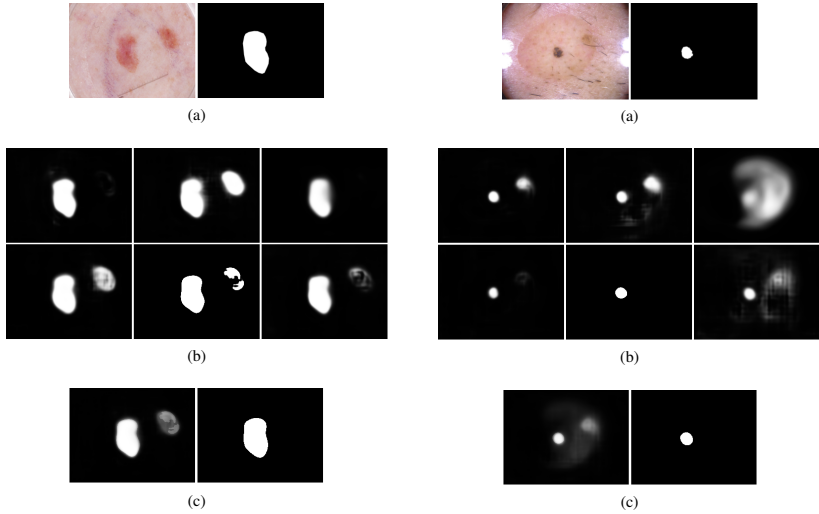


Figure 3.6: (a) Input images and their ground truths. (b) The output prediction of our CDNNs. Top row shows, from left to right, the output of CDNN₁, CDNN₂ and CDNN₃. Bottom row shows CDNN₀, CDNN₅ and CDNN₆. (c) Outputs ensemble before and after binarization.

mentation. This can be explained with the resizing process applied to every synthetic image during training, which halves image dimensions along each axis smoothing GAN-generated artifacts discussed in Section 3.2.3 and highlighted in Fig. 3.3.

The loss designed in Eq. 3.3 rewards more marked predictions, discouraging values too distant from both 0 and 1 in the output image. This behaviour amplifies the impact of CDNN₅ when it is employed in an ensemble method, thus reducing the overall accuracy. Hence, to obtain the final score, we also remove CDNN₅ and apply the ensemble strategy on the output results of best performing networks (*i.e.* networks corresponding to the bold values in Table 3.2), obtaining a final jaccard index of **0.789**, which is above the state-of-the-art accuracy of 0.782 [67] and the challenge winner result of 0.765³.

³2017 challenge scoreboard at <https://bit.ly/2yUhs30>

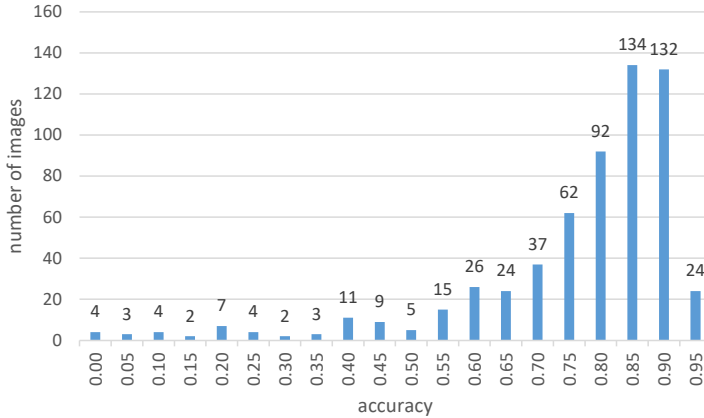


Figure 3.7: Histogram of Jaccard Index values for individual images from our top performing architecture.

As stated in [77], segmentation masks that achieve a Jaccard Index of 0.8 or above tend to appear visually correct, whereas when the accuracy falls under a threshold somewhere between 0.6 and 0.7, the correctness of the segmentation can be debated. The histogram in Fig. 3.7 illustrates the accuracy obtained by our architecture on each skin lesion image in the test set. We obtain a Jaccard Index below 0.6 for 69 images, and below 0.7 for 119, suggesting a failure rate between 11% and 20% on the 600 samples test set.

3.4.3 About Scalability

The huge cost of well-annotated data for medical deep-learning-based approaches is a well known problem. The proposed method can serve as a great tool to obtain cheap medical images in many different fields.

To explore the whole effectiveness of the described framework, we extrapolate three different subsets of the ISIC dataset, respectively composed by 500, 1000 and 1500 images. For each subset, a DCGAN is trained and then employed in the two-step training process of our baseline network, CDNN₀.

Training a GAN with a small set of images is not a trivial task: quality

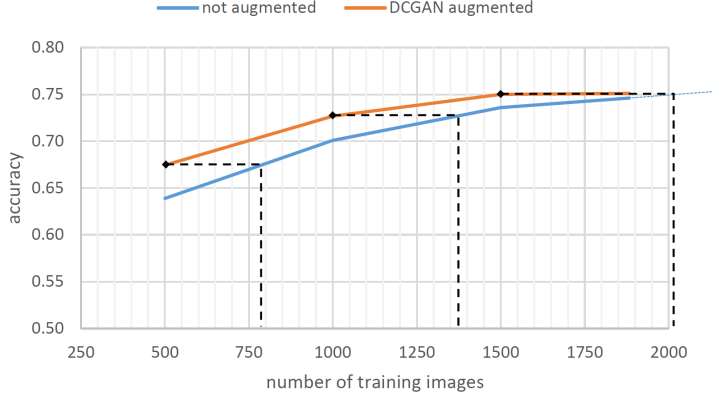


Figure 3.8: Accuracy obtained on the test set by CDNN_0 , trained employing reduced subsets of the dataset. The chart analyzes the effectiveness of the proposed augmentation process (using DCGAN) with respect to the availability of a larger dataset.

and diversity of the generated images decrease with the size of the training set. However, given a small amount of original annotated samples, synthetic images can improve the segmentation accuracy despite their low overall quality, as shown in Fig. 3.8. The chart reveals that, given an early stage project with a limited dataset, one would need up to 58% more annotated images to obtain the improvement given by our proposal. Moreover, it is clear how the complexity of the problem adds a lot of relevance to every apparently small accuracy increase.

Despite the lower results obtained on CDNN_0 (Table 3.2), we extend the scalability experiment to the LAPGAN. Fig. 3.9 displays how, on a dataset constituted by only 500 annotated images, LAPGAN-generated samples widely outperform their DCGAN-generated counterpart. As the number of annotated images rises, so do the performances of the DCGAN, suggesting that the DCGAN actually requires to be trained with more annotated images in order to produce data useful for our cause. When evaluating the method over a 500 images dataset, a LAPGAN can be employed to obtain an accuracy boost equivalent to 138% more real annotated images.

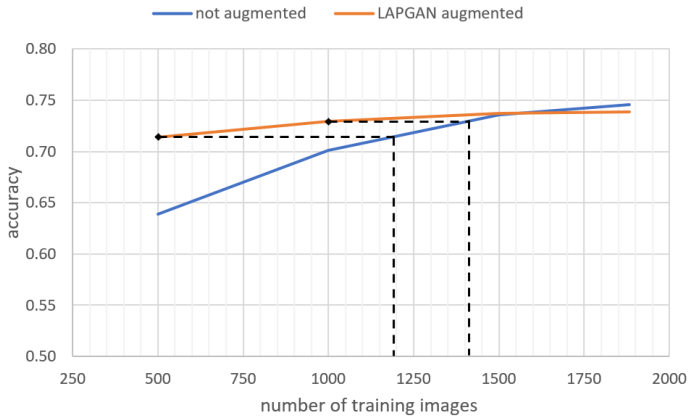


Figure 3.9: Accuracy obtained on the test set by CDNN_0 , trained employing reduced subsets of the dataset. The chart analyzes the effectiveness of the proposed augmentation process (using LAPGAN) with respect to the availability of a larger dataset.

3.5 Conclusion

This chapter proposed a new method to exploit GANs in the data augmentation process, mitigating the need of medical manually annotated data, which are very expensive to obtain. In order to improve results on the skin lesion segmentation task, GANs are used to generate both skin lesion images and their segmentation masks, creating a tool for *automatic data augmentation* that can be integrated in any supervised learning model. An improved version of the state-of-the-art architecture for automated skin lesion segmentation has been also developed and described in the chapter. We designed a two-step training process to exploit synthetic data, and compared the impact that two different types of state-of-the-art GANs can have on the final accuracy, obtaining results above the state-of-the-art. Finally, a stress test revealed the effectiveness that the proposed framework could have on similar tasks with a restricted dataset, revealing a winning trade-off between the quality of generated images and the segmentation accuracy.

Chapter 4

Skin Lesion Segmentation Ensemble with Diverse Training Strategies

4.1 Introduction

In this chapter, we address the problem of skin lesion image segmentation taking into account the trade-off between training a model from scratch and employing images of a very different nature through transfer learning. In particular, we design a CNN-based ensemble model that, exploiting different pre-training strategies, obtains the state-of-the-art performance on skin lesion segmentation when trained on the ISIC skin lesions dataset and on synthetically generated samples.

The rest of the chapter is organized as follows. Section 4.2 describes the pre-training strategies adopted to push the effectiveness of the proposed ensemble strategy, which is then described in Section 4.3. In Section 4.4 the performance of our proposal are compared with state-of-the-art model, highlighting its validity. Finally, Section 4.5 draws some conclusions.

4.2 Learning to Extract Features

In order to design an ensemble of CNNs that can effectively segment skin lesion images, we select DeepLabv3+ [30] with a ResNet-101 [5] backbone as the baseline architecture, and explore several pre-training strategies. This Section discloses three different pre-training methods employed to obtain three different initializations of ResNet-101, the features extractor for our segmentation neural network.

4.2.1 ISIC Task 3

The International Skin Imaging Collaboration, in 2018, hosted a challenge split in three tasks: lesion segmentation, lesion attribute detection and disease classification. A public dataset of 10 015 dermoscopic images was divided by experts into 7 different classes, illustrated in Fig. 4.1. In the interests of obtaining a features extractor effective on dermoscopic images, we train ResNet-101 to correctly classify skin lesion images in the 7 original classes, taking advantage of the 10 015 annotated samples from the ISIC dataset [80]. The network is trained for 15 epochs, using a weighted Cross-Entropy Loss, an initial learning rate of 0.0001, and the Adam optimizer [76]. Fig 4.2 provides a useful visualization of which sections of the input image mostly affect the hidden representation that the network obtains after the last convolutional layer, right before employing the fully-connected layers designed for classification. The heatmaps obtained by applying the Grad-CAM technique [65], suggest that our features extractor implicitly learns to coarsely detect the skin lesion.

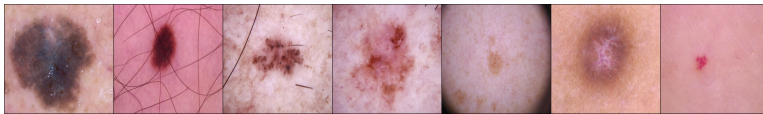


Figure 4.1: One sample of each class from the third ISIC challenge task. From left to right: Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis, Benign keratosis, Dermatofibroma, and Vascular lesion.

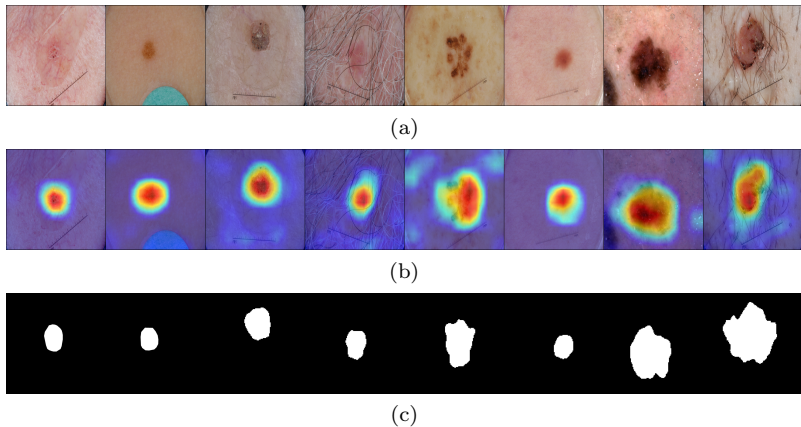


Figure 4.2: Samples from the 2017 ISIC validation set (a), with corresponding Grad-CAM heatmaps (b) and segmentation masks (c).

4.2.2 Autoencoder

Data annotation for the image segmentation task is extremely expensive, it requires an expert dermatologist to perform an extremely time consuming and uninspiring job. Therefore, we try to exploit many dermoscopic images that were never given manual annotation through a form of unsupervised learning. Taking advantage of the encoder-decoder architecture of DeepLabv3+, an autoencoder is built in order to obtain a version of ResNet able to map skin lesion images to a meaningful hidden representation. The model of our autoencoder is identical to DeepLabv3+ [30], with 3 output classes (RGB channels). It is trained using, again, the 10 015 images from the ISIC classification task for 10 epochs, employing the mean squared error loss. The learning rate is initially set to 0.0001 and then influenced by the Adam optimizer [76]. Fig. 4.3 shows that the autoencoder struggles to recreate the original background skin color and texture, but promptly generates a coherent lesion in the correct position, which is the most important element for our final goal.

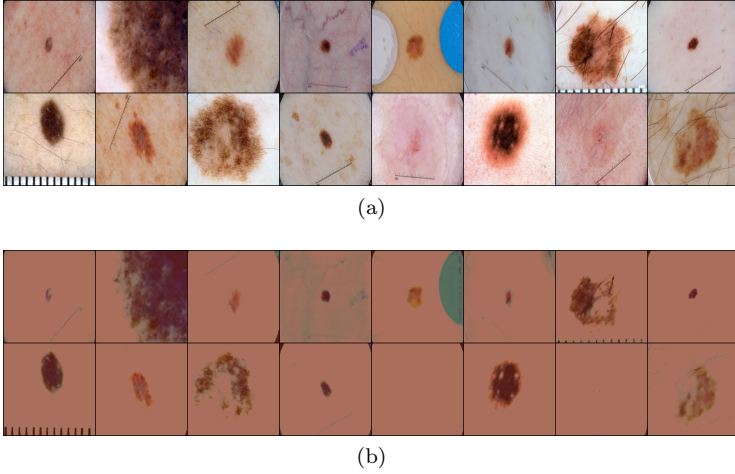


Figure 4.3: Samples from the 2017 ISIC validation set (a) and the corresponding autoencoder output (b).

4.2.3 GAN

Generative Adversarial Networks (GANs) are often used to create unlabeled examples, which cannot be directly employed for the training of a supervised algorithm [70]. Following the approach described in detail in Section 3.2, we modify the GAN proposed by Karras *et al.* [81] in order to feed it 4-channels images: the first three channels are the R, G and B components and the fourth one is the binary segmentation mask. Instead of generating a fixed number of samples, the generator is required to provide new couples image-mask for each training batch. An example of the generated images is shown in Fig. 4.4.

The GAN is trained to generate 256×256 images and segmentation masks. Most images present realistic details, like a well delivered presence of hair, black corners representing real camera characteristics, and pen marks. Moreover, the generative model produces segmentation masks that look very coherent with the respective generated images.

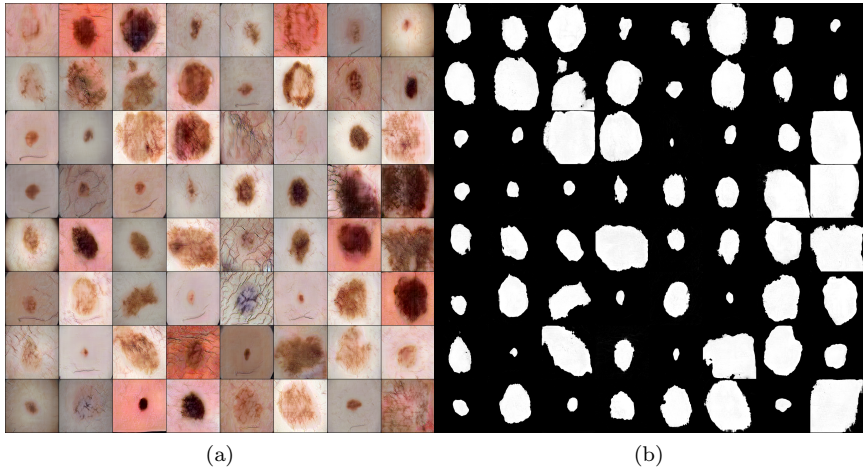


Figure 4.4: GAN-generated skin lesion samples (a) and their segmentation masks (b).

4.3 Learning to Segment Skin Lesions

To take advantage of the pre-trained backbones, DeepLabv3+ is fine-tuned following the training protocol outlined in [30]. Input images are resized to 513×513 pixels, the learning rate is initially set to 0.007 and then multiplied by $(1 - \frac{iter}{max_iter})^{0.9}$, following the poly policy [82]. The output stride is fixed to 16, the network is trained for 60 epochs, and the early stopping strategy is employed. During the whole fine-tuning process, the learning rate is multiplied by a factor of 10 for layers outside of ResNet-101, in order to take full advantage of the pre-training process applied to the features extractor. Data augmentation is performed by randomly rotating, flipping, shifting, shearing and scaling input images, and by changing the color contrasts.

To enhance the fine-tuning process, the official 2017 training set of 2000 images is enlarged using several dermoscopic images gathered from the public ISIC archive. In order to remove potentially disadvantageous samples, every supplementary image is first fed to a state-of-the-art segmentation CNN. Images segmented with an Intersection over Union lower

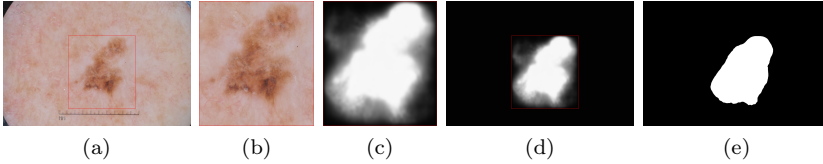


Figure 4.5: Detection-Segmentation Pipeline: sample image from the 2017 ISIC test set with its predicted bounding box overlapped (a), the cropped and resized sample (b), the output of the ensemble method (c) that is rescaled to the original size (d) and the final segmentation mask (e).

than 0.60 are considered incorrect and thus removed from further analysis. To build the final training set, we select the 1 500 dermoscopic images that offer a ground truth segmentation mask where the skin lesion fills at least the 10% of the image. We thus obtain a 3 500 images training set.

4.3.1 Detection-Segmentation Pipeline

In order to encourage the network to focus on the correct section of input images, we remove redundant background from the input samples. During training, every ground truth segmentation mask is exploited to obtain the bounding box containing the skin lesion, and every dermoscopic image is cropped accordingly (Fig. 4.5a and Fig. 4.5b). During inference, input samples are cropped in accordance with bounding boxes obtained through Mask R-CNN [31], pre-trained on the COCO dataset [83] and then fine-tuned on the 2017 ISIC original training set. Mask R-CNN produces bounding boxes with an Intersection over Union of 0.886 on the 2017 public test set. This pipeline, illustrated in Fig. 4.5, improves the Intersection over Union accuracy of the segmentation framework by 0.045. However, cropping images by means of an object detection network can erase meaningful portions of the segmentation CNN input, propagating the error from the first network to the next one. To tackle this drawback, bounding boxes can be directly supplied to the segmentation CNN, without forcing it to discard the sections of the input image classified as background by the detection tool. We thus feed the segmentation network with a 4-channels input: R, G, B, and the skin lesion bounding box in the form of a binary mask, where white pixels represent the foreground and

Table 4.1: Detection-Segmentation Pipelines Comparison.

Pipeline Strategy	Validation IoU
None	0.800
Crop	0.845
4-channels	0.848

black pixels represent the background. Improvements delivered by adding an object detection network to the process are shown in Table 4.1.

4.3.2 The Ensemble Method

As observed by Krogh *et al.* [84] and Chandra *et al.* [85], a key point on ensemble learning is that the hypotheses should be as accurate and as diverse as possible. In our work, such properties were prompted by applying different strategies. Even though the performance of the various CNNs are comparable, different networks focus on different features, greatly increasing the effectiveness of the ensemble. Fig. 4.6 displays how the features extractor pre-trained to build the autoencoder described in Section 4.2.2, seems to give much importance to darker shades of colours, whereas the CNN that employs the features extractor formerly used to classify natural images from ImageNet, produces segmentation masks less attentive to this particular feature, focusing on higher level structures instead.

For each pixel, its probability of being part of the skin lesion is obtained as the mean value across the selected CNNs. The output is then binarized with a dual-threshold method. A high threshold (0.80) is followed by blob analysis [78, 79] and the biggest object center is assumed to be the tumor center. Afterwards, a lower threshold (0.40) is applied and the final segmentation mask is given by the region which contains the tumor center. Whenever the first high threshold does not yield any object, we only apply the second one and keep its result as segmentation mask.

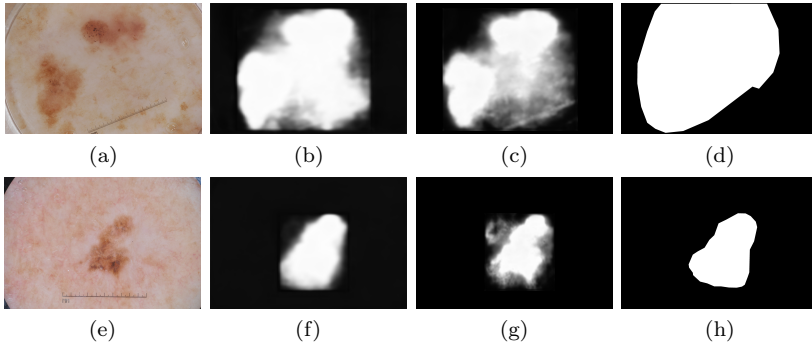


Figure 4.6: Examples of model outputs when using different pre-trained feature extractors. (a) and (e) are the original images, (b) and (f) are the outputs when using a features extractor pre-trained with ImageNet, (c) and (g) are the outputs of the CNNs that exploit the features extractor described in Section 4.2.2. Finally, (d) and (h) represent the ground truth.

4.4 Experimental Results

Experimental results of the proposed networks are summed up in Table 4.2. The first column shows the detection-segmentation pipeline of each CNN. It is important to notice that, when using the 4-channels strategy described in Section 4.3.1, the first layer of the features extractor is required to deal with four channels instead of three, and thus to be re-trained from scratch. Since every network described in Section 4.2 was pre-trained using dermoscopic images, we fine-tune them using the cropping strategy to avoid a random re-initialization of the first layer of filters.

In order to further increase the diversity between predicted masks, and hence the effectiveness of the ensemble architecture, we employ two different loss functions during training (second column of Table 4.2): the cross-entropy loss¹ in its binary form and the Tanimoto distance defined

¹Cross-Entropy is the standard loss function employed when training DeepLab [29].

Table 4.2: Analysis of the Neural Networks trained for the task. IoU is the Intersection over Union, TIoU is the Threshold Intersection over Union. The employed losses are Cross-Entropy (CE) and the Tanimoto Distance (Tan).

Pipeline Strategy	Loss	Pre-Training	Val IoU	Val TIoU	Test IoU	Test TIoU
Crop	CE	Classes	0.851	0.819	0.841	0.809
	CE	AE	0.856	0.837	0.838	0.810
	CE	GAN	0.854	0.834	0.838	0.814
	Tan	Classes	0.855	0.832	0.845	0.818
	Tan	AE	0.857	0.838	0.844	0.816
	Tan	GAN	0.850	0.831	0.841	0.817
4-channels	CE	None	0.847	0.816	0.834	0.805
	CE	ImageNet	0.848	0.823	0.837	0.808
	Tan	None	0.846	0.819	0.831	0.802
	Tan	ImageNet	0.859	0.840	0.845	0.819

as

$$L = 1 - \frac{\sum_{i,j} t_{ij} p_{ij}}{\sum_{i,j} t_{ij}^2 + \sum_{i,j} p_{ij}^2 - \sum_{i,j} t_{ij} p_{ij}} \quad (4.1)$$

where t_{ij} is the target value of the pixel at coordinates (i, j) , and p_{ij} is the real output. Note that t_{ij} is either 0 or 1, while p_{ij} is a real number in range $[0, 1]$.

The third column of Table 4.2 presents the pre-training procedure applied to each network. Values *Classes*, *AE*, and *GAN* refer to Sections 4.2.1, 4.2.2, and 4.2.3 respectively. The value *None* represents a network trained from scratch, included for comparison. The last four columns of the Table present the accuracy of each network, on both the validation set and the test set from the ISIC 2017 challenge. IoU stands for Intersection over Union, which is the official evaluation metric of the 2017 ISIC challenge. TIoU means Threshold Intersection over Union, an adaptation of Intersection over Union that reflects the number of images in which automated segmentation fails, by giving a score of 0 to images segmented with an Intersection over Union lower than 0.65. Both of the metrics show similar results when testing different initialization methods, but stress the im-

Table 4.3: Performance of the proposed method compared to the state-of-the-art. All the models have been trained and tested on the 2017 ISIC dataset. Reimplemented/retrained models are identified by *. IoU is the Intersection over Union, TIoU is the Threshold Intersection over Union.

Method	Test IoU	Test TIoU
Ours (ensemble)	0.850	0.827
SegAN [86]	0.785	—
GAN Augmented [67]	0.781	—
DCL-PSI [87]	0.777	—
DeepLabv3+* [30]	0.769	—
(RE)-DS-U-ResnetFCN34 [88]	0.772	—
SegNet* [89]	0.767	—
Challenge winners [15]	0.765	—
Tiramisu* [90]	0.765	—
U-Net* [28]	0.740	—

portance of employing a pre-trained feature extractor: CNNs trained from scratch (value *None* in the third column) show the worst results. Table 4.3 displays the results obtained by merging the output of resulting CNNs through the ensemble strategy described in Section 4.3.2. The proposed pipeline clearly outperforms state-of-the-art segmentation algorithms when applied on skin lesions.

4.5 Conclusion

With this chapter, we tackled the problem of skin lesion segmentation by a different perspective, introducing a novel ensemble strategy to improve state-of-the-art results. We presented multiple ways to initialize a features extractor without the need to employ biases-inducing datasets, and highlighted the relevance of employing transfer learning in dermoscopic image segmentation. Moreover, we identified skin lesion detection as an extremely valuable preliminary step for skin lesion segmentation. As a matter of fact, both cropping the lesion bounding box and feeding it as a fourth image channel yield a major improvement to the overall semantic segmentation

performance. The designed segmentation pipeline makes use of the findings described in the previous chapter, and takes advantage of multiple additional pre-training methods to improve the overall performance. Experimental results on the ISIC 2017 dataset show the effectiveness of our approach on skin lesion segmentation. The source code of the proposed model is available in [91].

Chapter 5

A Deep Analysis on High Resolution Dermoscopic Images Classification

5.1 Introduction

A bigger neural network size has always been considered a synonym for better accuracy, since a larger amount of parameters and layers means a greater capability to learn important filters, and therefore to capture meaningful features within an image as long as its resolution is high enough. However, deeper networks are more prone to overfitting and harder to regularize during training, in addition to requiring a considerable amount of time to be trained [92]. This often produces inefficient architectures, which do not improve results yielded by faster, shallower networks. Tan *et al.* [6] propose an efficient way to scale up convolutional neural networks width, depth, and image resolution by performing exhaustive experiments on several datasets of natural images.

However, medical images present several dissimilarities from natural ones, and thus require an individual analysis [93]. As a matter of fact, in dermoscopic images, the difference between background (human skin) and foreground (skin lesion) can be less visible than in most other scenarios. Indeed, sharpening filters are broadly used in this field to enhance lesion

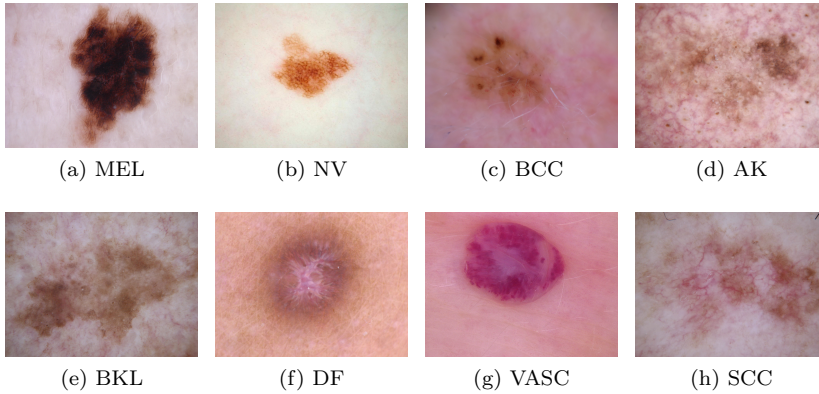


Figure 5.1: Samples of the 2019 ISIC dataset. (a) Melanoma - MEL, (b) Melanocytic Nevus - NV, (c) Basal Cell Carcinoma - BCC, (d) Actinic Keratosis - AK, (e) Benign Keratosis - BKL, (f) Dermatofibroma - DF, (g) Vascular Lesion - VASC, (h) Squamous Cell Carcinoma - SCC.

borders. Moreover, dermoscopic images present numerous artifacts such as black round borders, pen drawings, rulers, and hair, which must be ignored when seeking meaningful patterns within an image.

This particular domain of medical imaging is characterized by a very high resolution, which must be taken into account when aiming for good classification accuracy. Additionally, a huge quantity of data augmentation strategies can be performed without altering the nature of the skin lesion, and this can be exploited during inference by merging the outputs of models that are robust against different combinations of simple transformations.

Finally, it is crucial to make use of calibrated models for a correct behavior in critical decision scenarios, such as medical diagnosis, in which the ultimate goal is not substituting an expert practitioner, but providing a reliable measure of our degree of uncertainty in order to assist the final decision.

In this chapter, we perform a thorough investigation about the performance of state-of-the-art architectures for natural images classification when applied to dermoscopic image analysis. A comprehensive discussion on how the major hyperparameters (the size of the network, image resolution, data

augmentation process, amount of available data, and model calibration) affect neural network capabilities is provided, and a novel ensemble method for dermoscopic image classification is designed. The proposed ensemble yields a balanced accuracy of 0.593 on the official 2019 ISIC challenge, achieving the third best result. However, further experimental results show that our approach outperforms the winners of the challenge when the two algorithms are trained and tested using the same data.

The rest of this chapter is organized as follows. Section 5.2 introduces the 2019 ISIC dataset and the proposed preprocessing pipeline. The designed ensemble architecture is presented and motivated in Section 5.3, and evaluated in Section 5.4 through an in-depth analysis. Finally, in Section 5.5 conclusions are drawn.

5.2 Dermoscopic Images

Since 2016, the International Skin Imaging Collaboration (ISIC) has begun to aggregate a large-scale, publicly available dataset of dermoscopic skin lesions images (Fig. 5.1) and hosting multiple challenges and workshops [77]. The availability of this substantial amount of dermoscopic images allowed to significantly improve the performance of machine learning algorithms. This dataset, also known as the ISIC archive, is designed both for clinical training and to support research toward automated skin cancer analysis.

The 2019 version of the ISIC archive contains a total amount of 25 331 dermoscopy labeled images, belonging to eight different classes (*i.e.* types of skin lesion) [94]. Images have been collected in several years, from different centers, and using multiple devices. For this reason, their resolution ranges from 450×600 to 1024×1024 pixels.

The available data is heavily imbalanced, as samples are distributed among classes as follows:

1. Melanoma (MEL) - 17.8%
2. Melanocytic Nevus (NV) - 50.8%
3. Basal Cell Carcinoma (BCC) - 13%
4. Actinic Keratosis (AK) - 3%
5. Benign Keratosis (BKL) - 10%

6. Dermatofibroma (DF) - 0.9%
7. Vascular Lesion (VASC) - 1%
8. Squamous Cell Carcinoma (SCC) - 2.4%

However, the official 2019 test dataset counts an additional class, which is not available in the training partition of the data. This class, named *none of the others*, contains dermoscopic images of different natures that do not belong to any of the other eight classes. In order to correctly evaluate such a heavily imbalanced task, the 2019 official challenge judges the participants by means of the Balanced Accuracy metric, which is computed as the average sensitivity among classes. This metric gives the same importance to each class, regardless of how much it is represented in the test set. Additional metrics such as the Area Under the ROC Curve (AUC) are presented in the official leaderboard, but not employed in the final scoring.

5.2.1 Preprocessing

Dermoscopic images present several dissimilarities from natural images such as a very high resolution, low color variability within an image, and many unnatural artifacts like pen marks or black corners introduced by acquisition devices. This is mainly because the subject of these images is human skin, and because of the particular acquisition technique that aims to manipulate how the light hits the epidermis. The discrepancy can be noticed by just computing the dataset statistics and using Gaussian distributions as an approximation of the real pixels values distribution (Fig. 5.2). It is thus crucial to carefully choose preprocessing steps and data augmentation strategies, instead of reusing procedures developed for natural images. Hence, a specific dataset mean and standard deviation must be computed and exploited for input normalization, which is an essential step to obtain an efficient training process. In order to ensure good results, the same values must be used to perform input normalization during inference.

Four main characteristics are widely recognized as primary attributes for the detection of melanocytic lesions, which are asymmetry, border irregularity, color variegation, and a diameter greater than 6 millimeters [95]. The analysis of these four features is also known as the ABCD rule, and it represents the basic guideline to preserve semantic information within

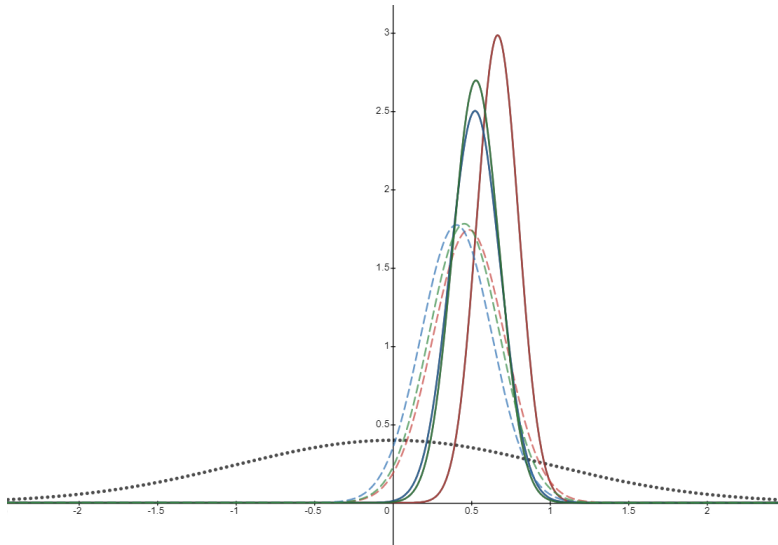


Figure 5.2: Approximation of the real pixel values distribution through Gaussian distributions. Dashed, lighter lines represent the distribution of channels R, G, and B of the ImageNet Dataset, whereas solid, darker lines represent the ISIC dataset. The grey dotted line is the Gaussian distribution with mean 0 and standard deviation 1, which we aim to obtain after input normalization.

dermoscopic images. Since image sizes are not constant, the first step is to obtain a dataset of squared images by replicating the border of rectangular pictures along the shorter side, in order to not change the shape nor the size of skin lesions, which are key factors for the diagnosis. The next step concerns Data Augmentation (DA), an important and well-known operation that can be performed during training to improve the effectiveness of neural networks [96]. This process consists in generating new data items by applying very simple transformations to existing training samples, without changing their semantic content. This technique aims to improve the robustness of an algorithm against used transformations and thus boost the final accuracy [67, 97]. With respect to natural images, dermoscopic ones can benefit from a larger amount of data augmentation steps, consider-

ing that transformations like vertical flips or rotations do not alter their semantic content. Furthermore, dermoscopic images are often refined by means of sharpening filters to emphasize lesion borders and make images easier to inspect for expert dermatologists. This processing technique can however lead neural networks to erroneously focus on low-level features that are trivial. To avoid this drawback, images can be randomly blurred through gaussian kernels, teaching CNNs to ignore differences caused by the use of diverse sharpening filters. Fig. 5.3 shows an example image before and after the employment of the blurring filter; the enlarged version of the images exposes that low-level artifacts introduced by the sharpening filters get mitigated thanks to the Gaussian filters. Blurring images with random values allows trained CNNs to increase their steadiness against a wide spectrum of sharpening filters that are used during acquisition.

As previously mentioned, randomly flipping and rotating images during training never change the semantic content, and always yields good results. Moreover, the manipulation of contrast, hue, and saturation are very common techniques for augmenting this kind of images. This is because of the assumption that different acquisition devices can alter the representation of similar colors, in addition to the effect that different light settings and natural skin tones can have on images captured with the same camera. Even though we find accuracy gains yielded by these data augmentation strategies to often be minor, they can all be exploited to increase the robustness of the inference method described in the following Section.

Finally, the network training regularization can be improved through the CutOut strategy [98], and by adding Poisson distributed random noise to the input image. Fig. 5.4 presents two samples of dermoscopic images, and their appearance after applying the random data augmentation described in the first row of Table 5.1, the effects of the cutout method are not displayed as it is applied after pixel normalization.

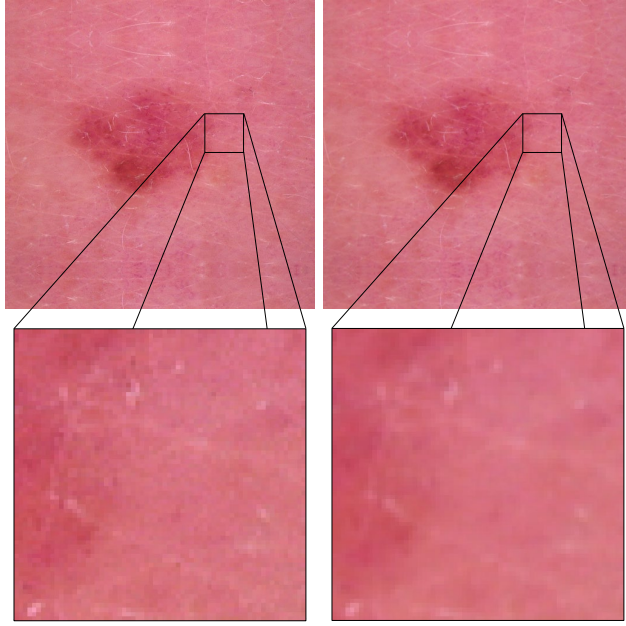


Figure 5.3: On the left, the original 512×512 input image and an enlarged detail. On the right, the same image after a blurring filter is applied, with the same enlarged patch. The original image presents a noisy checkerboard effect, introduced by a sharpening filter. This effect is greatly reduced in the picture on the right.

Table 5.1: Different data augmentation configurations. When not specified, configuration 0 is employed.

Conf	Flips	Rot	Gauss	Cutout	Noise	H&S	Contrast
0	✓	✓	✓	✓	✓	✓	✗
1	✓	✗	✗	✗	✗	✗	✗
2	✓	✓	✓	✓	✗	✗	✗
3	✓	✓	✓	✓	✓	✓	✓

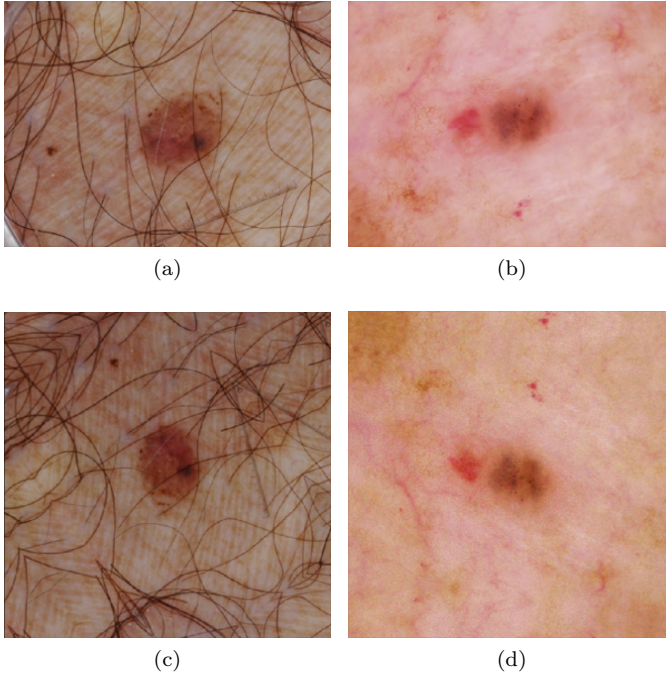


Figure 5.4: Random samples of dermoscopic images (top), coupled with the results of random data augmentation (bottom) performed by flipping and rotating the images, applying Gaussian filters, adding noise with a Poisson distribution, and manipulating hue, and saturation.

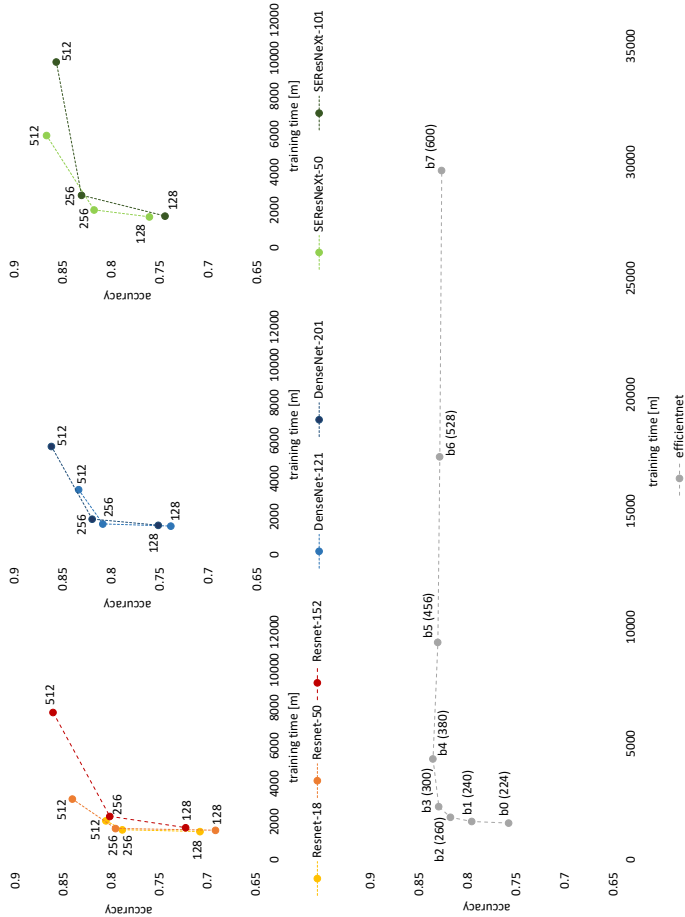


Figure 5.5: Correlation between **balanced accuracy** (y axis) and **training times** in minutes (x axis) for different network architectures. Within charts, each point represent one trained model, characterized by the network architecture and the input image size. EfficientNet CNNs are trained with their conventional input image resolution, whereas every other network is trained with three different image sizes: 512×512 , 256×256 and 128×128 . Lower training times are required for the smallest resolution, and viceversa.

5.3 High Accuracy Through Growth

In order to improve classification accuracy, convolutional neural networks can be scaled up in multiple ways. The most common approaches can be summarized as increasing either the number or the size of convolutional filters within a model. In addition to the growth of time required to complete the training process, scaled up networks necessitate a more careful regularization, and are more prone to overfitting [92]. Both of these problems are especially relevant when the volume of available training data is not large enough. On the other hand, increasing the input image resolution also boosts CNN performance regardless of the network architecture employed, at the cost of incrementing the required training time. Adopting a larger input size makes it possible to fully exploit the high resolution that characterizes dermoscopic images. As a matter of fact, convolutional neural networks can be effectively fine-tuned by using images of a different resolution than the one used during the pre-training process, Table 5.2, Table 5.3, Table 5.4 and Table 5.5 show how increasing resolution is the most effective way to boost the skin lesion classification accuracy for every tested architecture (EfficientNet [6], ResNet [5], DenseNet [43], and SEResNeXt [99]). As depicted in Fig. 5.5, growing the input size from 256 to 512 yields improvements showcased by balanced accuracy boosts that range from 1.7% to 5.9%, whereas growing the network size by increasing the number of parameters produce smaller boosts and, in some cases, even drops in accuracy.

5.3.1 Probabilistic Model

In medical contexts, it is crucial to provide both good discriminative power and reliable confidence. Therefore, in this Section we describe our probabilistic model, which is divided in how we assign probabilities, and how we combine them.

Given a set of independent and identically distributed (i.i.d.) labeled pairs of samples $\mathcal{O} = \{x_i, t_i\}_{i=1}^N$ made up from images x_i with their corresponding categorical class labels t_i , we estimate the joint distribution of a set of M models, also known as the model ensemble, parameterized by $\Theta = \{\theta_m\}_{m=1}^M$.

Given the model parameters θ , assuming independence between the models and a non-informative prior over θ , the joint distribution factorizes

Table 5.2: Training times (expressed in minutes), balanced accuracy, and area under the roc curve (AUC) of several neural networks architectures with 512×512 input size.

Net	Training Time	Balanced Accuracy	AUC
DenseNet-201	5700	0.862	0.980
DenseNet-121	3420	0.834	0.972
SEResNeXt-101	9804	0.857	0.981
SEResNeXt-50	5928	0.867	0.982
ResNet-18	2052	0.806	0.975
ResNet-50	3192	0.841	0.977
ResNet-152	7752	0.861	0.978

as:

$$p(t^1, t^2, \dots, t^M, \Theta | \mathcal{O}) = \prod_m p(t^m | \mathcal{O}, \theta_m) \cdot p(\theta_m) \quad (5.1)$$

Learning involves maximizing this joint model, which under our assumptions is the same as learning each of the conditional distributions $p(t^m | \mathcal{O}, \theta_m)$ separately, by optimizing the cross-entropy loss without any form of regularization. The proposed method maps each of these conditional distributions with different network architectures and data augmentation techniques. The optimal number of networks and preferred model architectures are chosen using a validation set, because using the evidence framework [100] for automatic Occam’s razor is intractable, and approximations are a computational burden. The final probabilistic vector assigned to a given test sample x^* is obtained by computing the posterior distribution of the label, given the models $p(t^* | t^1, t^2, \dots, t^M, x^*)$. This posterior can be computed in several ways: with standard model average [101], with a learned combination, with boosting techniques [102] or Bayesian classifiers [103], *i.e.*, substituting the weights by the posterior distribution of the models given the data.

Considering standard model average, for a test sample x^* we assign the label t^* with confidence p^* as follows:

Table 5.3: Training times (expressed in minutes), balanced accuracy, and area under the roc curve (AUC) of several neural networks architectures with 256×256 input size.

Net	Training Time	Balanced Accuracy	AUC
DenseNet-201	1870	0.820	0.975
DenseNet-121	1619	0.809	0.966
SEResNeXt-101	2782	0.831	0.978
SEResNeXt-50	2006	0.818	0.975
ResNet-18	1573	0.789	0.968
ResNet-50	1641	0.796	0.965
ResNet-152	2280	0.802	0.972

$$\begin{aligned}
 p(t^*|x^*, t^1, t^2, \dots t^M) &= \frac{1}{M} \sum_m p(t^m|x^*, \hat{\theta}_m) \\
 \hat{t}_i &= \operatorname{argmax} p(t^*|x^*, t^1, t^2, \dots t^M) \\
 \hat{p}_i &= p(t^* = \hat{t}_i|x^*, t^1, t^2, \dots t^M)
 \end{aligned} \tag{5.2}$$

Moreover, we augment this posterior probability with a small set of models for which we perform data augmentation at inference time, *i.e.* we combine the predictions of modified versions of a given test sample x^* through image transformations, as described in Section 5.3.2.

However, one of the consequences of doing Maximum A posterior Probability (MAP) estimation is that, when dealing with unbalanced datasets, many local optima tend to ignore the unrepresented classes. A possible solution is to compute the predictive distribution under a Bayesian paradigm, but this is again impractical for our purposes. The solution we adopt is to turn the discriminative classifier into a generative one, and subtract the prior information over the classes:

$$p(t_1|x, \theta) \propto p(x|t_1, \theta) \cdot p(t_1|\theta) \tag{5.3}$$

Then, scaling the posterior probability by the inverse of the prior information $1/p(t_1|\theta)$ we force our network to learn $p(x|t_1, \theta)$. This can be

Table 5.4: Training times (expressed in minutes), balanced accuracy, and area under the roc curve (AUC) of several neural networks architectures with 128×128 input size.

Net	Training Time	Balanced Accuracy	AUC
DenseNet-201	1550	0.752	0.950
DenseNet-121	1505	0.739	0.942
SEResNeXt-101	1687	0.745	0.960
SEResNeXt-50	1642	0.761	0.958
ResNet-18	1482	0.708	0.937
ResNet-50	1550	0.692	0.930
ResNet-152	1687	0.723	0.946

viewed as learning the posterior probability $p(t|x)$ assuming equal prior distributions $p(t)$, thus, training the model not to discard the unrepresented class. The model evidence $p(x)$ plays no role in this re-scaling as it is common to all the classes, thus it can be absorbed in the learning rate. In practice, this can be efficiently done by a weighted cross-entropy loss [104].

Model Calibration.

In the medical diagnosis field the goal should not be to take an action on behalf of an expert practitioner, but rather to assist his/her choice [105]. We thus end up discussing model calibration, mandatory for optimal decision. In such scenario, the decision made by an expert practitioner is differently influenced if we provide a confidence of 0.9 over a confidence of 0.4. This means that the provided information will only be useful if it is reliable, and this is achieved by a proper calibration of the probabilistic model. For a wider description of model calibration in a classification scenario see [106] and references therein.

The intuitive motivation of having a calibrated model can be seen from a more theoretical perspective. Taking into account that our aim is to combine expert knowledge and probabilistic information in an optimal way, we can formalize the problem using the Bayes decision rule, where the selected action α_i is the one that minimizes Bayes risk:

Table 5.5: Scaled EfficientNet performance results. Training times are expressed in minutes.

Net	Image Size	Training Time	Balanced Accuracy	AUC
EfficientNet-b0	224	1573	0.758	0.961
EfficientNet-b1	240	1642	0.796	0.967
EfficientNet-b2	260	1824	0.818	0.971
EfficientNet-b3	300	2280	0.830	0.974
EfficientNet-b4	380	4332	0.836	0.978
EfficientNet-b5	456	9348	0.831	0.975
EfficientNet-b6	528	17328	0.829	0.968
EfficientNet-b7	600	29640	0.827	0.967

$$\begin{aligned}
 \mathbf{R}(\alpha_i|x) &= \sum_j \lambda_{ij} p(t_j|x) \\
 \alpha_i &= \underset{i}{\operatorname{argmax}} \mathbf{R}(\alpha_i|x)
 \end{aligned}
 \tag{5.4}$$

$\mathbf{R}(\alpha_i|x)$ denotes the Bayes risk and λ_{ij} is the loss incurred in deciding class i when the true value is j . Note that for the particular case in which $\lambda_{ii} = 1$ and $\lambda_{ij} = 0$ this rule ends up being the maximum a posteriori decision rule. Furthermore, expert knowledge can be incorporated in these coefficients.

It is well known that this rule guarantees optimal decision if the data reflects the distribution $p(t|x)$ [102]. In practice, these distributions are substituted with our model $p(t|x, \theta)$. This means that the lower the gap between the model and the data distribution, the closer we are to an optimal decision. In general, this is achieved by models both able to separate between classes (a property known as discrimination or refinement [107]), and able to assign correct probabilities based on how the data is distributed (calibration).

Motivated by the properties of calibrated models, we propose to ensemble neural networks that have been previously calibrated. Considering that ensembles aim to combine probabilistic information, one should expect that the combination of more reliable classifiers, as those with a

proper calibration, should provide more reliable final posterior probability (Equation 5.4). In fact, merging multiple classifiers usually boosts the accuracy, as the combination of different local minima is a better representation of the data distribution [108, 109]. Following this observation, it has been proved in [101] that the average combination of classifiers tends to also calibrate the final predictions. We extend this consideration using an ensemble of calibrated models.

5.3.2 Averaging Calibrated Probabilities

We observe that data augmentation can be employed also during inference to provide calibrated probabilities. As a matter of fact, a multitude of data augmentation strategies can be applied to dermoscopic images without altering their semantic content (Table 5.1). Although not all augmentation steps induce a performance boost, they can all be exploited during inference by feeding each test image multiple times to each network, and by subjecting each image to the same random data augmentation process performed during training. By averaging the predictions of a single network, we employ an ensemble technique that increases both the overall accuracy and calibration without the need of training additional networks. The only drawback is an inference time overhead. This method can be especially beneficial when multiple networks are trained with diverse data augmentation strategies and then merged together. In these cases, DA increases the ability of the framework to be robust to small transformation that are irrelevant towards the final diagnosis.

To take this one step further, we calibrate single models by means of Temperature Scaling [106], an easy to integrate calibration technique with good performance in image classification. Empirical results show that the Data Augmentation Ensemble (DAE) improves both the accuracy and the calibration of the final ensemble of CNNs, whereas the usage of Temperature Scaling, despite successfully lowering the Expected Calibration Error (ECE) of a single model, degrades the overall ensemble calibration while having a small impact on the accuracy. However, this results can not yet be generalized for more sophisticated calibration techniques given in [106, 110, 101, 111]. We illustrate the whole proposed pipeline in Fig. 5.6.

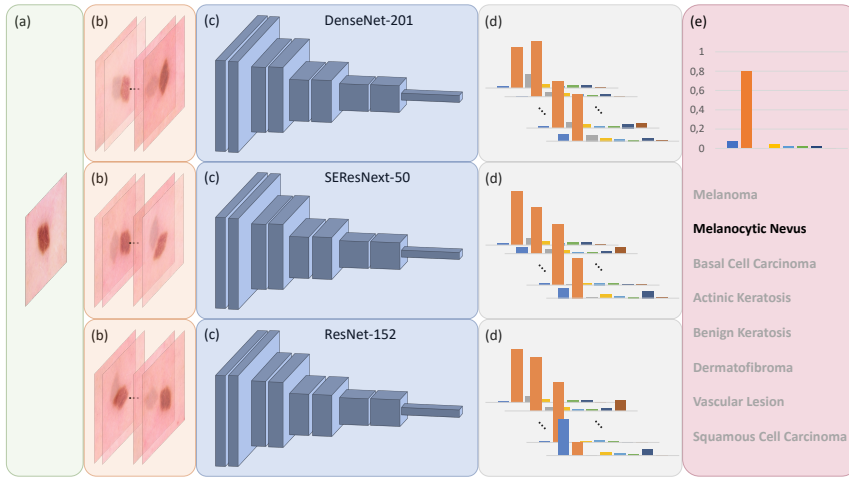


Figure 5.6: Illustration of the proposed ensemble method. Given an input image (a) at inference time, multiple versions of it are obtained by means of random data augmentation. Augmented images (b) are then fed to a CNN (c) and calibrated outputs (d) are obtained. This process is repeated over multiple CNNs and, finally, all of the network outputs are averaged together in the last step in order to obtain the final prediction (e).

5.4 Experimental results

This section presents the impact that different architectures, image resolutions, augmentation strategies, and dataset sizes have on the classification capabilities of several neural networks. Each network is trained using stochastic gradient descent (SGD) with momentum and a plateau learning rate scheduler. A validation set of 1000 images is used to monitor the accuracy of the model at each epoch and to apply the early stopping technique. Each network is tested on 5000 images and trained on a set of 19331 images, using a cross-entropy loss function which is weighted according to the inverse prior probability of each class. Ratios between classes are preserved in the training set as well as in both the validation and the test set. In order to avoid trivial comparisons, training times are computed on the same machine, equipped with one NVIDIA GTX 1080 Ti with 11 GB of

memory.

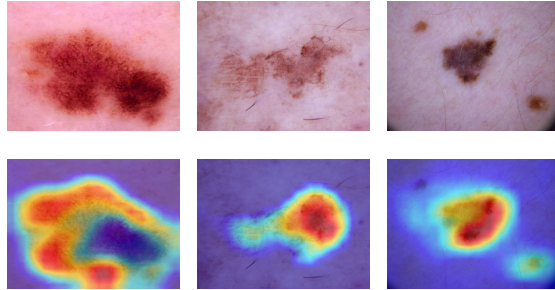


Figure 5.7: Samples from the 2019 ISIC dataset and attention heatmaps obtained by means of the Grad-CAM method. Besides always locating the skin lesion, the network can aim its attention at specific sections such as lesion borders (leftmost image) or darker patches (rightmost image).

In order to provide a qualitative visualization of the experimental results, we make use of the Grad-CAM method [65] to show the attention heatmaps associated to one of our models. Fig. 5.7 shows the specific part of the image on which the trained model focuses to make its prediction. On the other hand, quantitative results are expressed by means of two different metrics: Balanced Accuracy and Area Under the ROC Curve (AUC).

In Table 5.2, Table 5.3, Table 5.4 Table 5.5, and Fig. 5.5 we analyze the performance of different networks when image resolution, network depth, and network width are scaled up. The results of this investigation reveal that image resolution is the most relevant hyperparameter for this task. As a matter of fact, the accuracy obtained by EfficientNet-b5 can be achieved by smaller networks like ResNet-50, by virtue of merely increasing the size of input images. Indeed, deeper versions of EfficientNet are unable to take advantage of their width, depth, and image resolution without any supplementary data, yielding worst accuracy results than EfficientNet-b4.

Table 5.6 and Fig. 5.8 show the results obtained by merging three different networks, employing the ensemble technique described in the previous Sections and pictured in Fig. 5.6. For the sake of providing a fair evaluation of the proposed method, we make use of the guidelines and the code provided by the winners of the 2019 ISIC challenge (DAISY Labs [112]),

Table 5.6: Results obtained through a calibrated ensemble using different Data Augmentation configurations (DA) and fixed 512×512 image size over the training partition.

Net	Data Augmentation	Balanced Accuracy	AUC
DenseNet-201	2	0.861	0.980
SEResNeXt-50	0	0.867	0.982
ResNet-152	3	0.863	0.979
ensemble	-	0.888	0.988

		ground truth							
		MEL	NV	BCC	AK	BKL	DF	VASC	SCC
prediction	MEL	85	4	1	3	3	0	1	1
	NV	11	94	1	0	6	1	3	0
	BCC	1	1	95	6	1	1	1	9
	AK	1	0	1	84	3	1	0	7
	BKL	2	1	1	6	85	0	0	3
	DF	0	0	0	0	0	96	0	0
	VASC	0	0	0	0	0	0	95	0
	SCC	0	0	1	3	2	0	0	79

Figure 5.8: Confusion Matrix to show results of the calibrated ensemble detailed in Table 5.6, *i.e.* using different Data Augmentation configurations (DA) and fixed 512×512 image size over the training partition. Each cell contains the percentage of images of a specific class defined by the column, that were assigned to the class specified by the row. Accordingly, the main diagonal (green squares) display the architecture sensitivity for each class.

Table 5.7: Results obtained through an ensemble of the three top performing architectures described in [112]. The two input strategies employed in the paper are Random-Resize cropping (RR) and Same-Size cropping (SS).

Net	Input Strategy	Balanced Accuracy	AUC
EfficientNet-b4	RR	0.800	0.965
EfficientNet-b5	SS	0.777	0.957
EfficientNet-b6	SS	0.773	0.957
ensemble	-	0.871	0.983

and build an ensemble using the three top performing architectures described in the chapter. We implement the preprocessing strategy detailed by the authors using the European Computer Vision Library (ECVL) [113]. Each model is trained and tested using the same data partition described at the beginning of this Section, and the results are displayed in Table 5.7. The first three rows of Table 5.6 and Table 5.7 display the performance of neural networks when tested individually, through a single forward pass, and with no ensemble techniques applied at inference time. On the other hand, the last row of Table 5.7 (ensemble) presents the results obtained by merging the output of three CNNs after applying the two prediction strategies described in [112], both of which involve multiple forward passes for each network. The comparison between Table 5.6 and Table 5.7 clearly shows that the proposed method outperforms the winners of the ISIC 2019 challenge, when the two algorithms are trained and tested using the exact same data.

In order to investigate how our framework performance is affected by the amount of available data, we randomly build three smaller version of the ISIC dataset with respectively 10 000, 5 000, and 1 000 images, always preserving the ratio between different classes. Table 5.8 underlines that dataset size is a key factor for obtaining good performance, as reducing the amount of available data always worsen the final accuracy. Moreover, the impact of high resolution is only reduced for drastic configuration, with very shallow networks and extremely little data available. In Table 5.9 we also present the effects obtained by increasing the number of performed

Table 5.8: Balanced Accuracy (BA) and Area Under the ROC Curve (AUC) obtained by ResNet architectures with different Training Partition Sizes (TPS) and image resolutions.

Network	TPS	512 × 512		256 × 256		128 × 128	
		BA	AUC	BA	AUC	BA	AUC
ResNet-18	10 000	0.699	0.952	0.690	0.945	0.625	0.918
ResNet-18	5 000	0.610	0.907	0.617	0.913	0.560	0.893
ResNet-18	1 000	0.431	0.856	0.450	0.863	0.416	0.826
ResNet-50	10 000	0.750	0.959	0.722	0.943	0.614	0.914
ResNet-50	5 000	0.663	0.930	0.637	0.909	0.511	0.883
ResNet-50	1 000	0.456	0.865	0.464	0.842	0.413	0.806
ResNet-152	10 000	0.773	0.964	0.739	0.951	0.648	0.918
ResNet-152	5 000	0.689	0.936	0.633	0.925	0.572	0.889
ResNet-152	1 000	0.522	0.891	0.495	0.863	0.433	0.831

data augmentation steps during the training process, with datasets of various sizes.

Finally, Table 5.10 displays the results obtained on the official 2019 ISIC challenge by following the guidelines defined in this chapter. The official metric of the challenge is the balanced accuracy, AUC values are added for completeness. Metrics are computed over 9 classes as described in Section 5.2, the *none of the others* class is not considered by our approach because no labeled data is provided.

The proposed ensemble strategy is the best performing method when compared to algorithms that do not employ an out of distribution detection technique to handle the ninth class, and take advantage of no additional data.

Table 5.9: Balanced Accuracy (BA) and Area Under the ROC Curve (AUC) obtained by ResNet-50 with different Training Partition Sizes (TPS), Data Augmentation configurations (DA), and image resolutions. Indexes of DA are the same introduced in Table 5.1.

TPS	DA	512 × 512		256 × 256		128 × 128	
		BA	AUC	BA	AUC	BA	AUC
19 000	1	0.804	0.973	0.754	0.958	0.626	0.923
19 000	2	0.837	0.976	0.781	0.963	0.704	0.939
19 000	3	0.846	0.977	0.770	0.960	0.660	0.921
10 000	1	0.725	0.956	0.664	0.931	0.557	0.900
10 000	2	0.730	0.959	0.705	0.938	0.589	0.905
10 000	3	0.743	0.964	0.708	0.935	0.617	0.909
5 000	1	0.642	0.927	0.586	0.902	0.482	0.872
5 000	2	0.667	0.926	0.635	0.911	0.516	0.872
5 000	3	0.656	0.928	0.639	0.910	0.549	0.871
1 000	1	0.432	0.845	0.413	0.837	0.337	0.767
1 000	2	0.453	0.855	0.450	0.842	0.377	0.800
1 000	3	0.459	0.869	0.457	0.847	0.406	0.800

Table 5.10: Results obtained on the official 2019 ISIC Challenge [114], with a description of data used for every method. The last column indicates methods that employ OOD strategy.

Method	Balanced Accuracy	AUC	Employed Images	OOD
DAISYLabs [112]	0.636	0.923	ISIC 2019 + additional data	✗
DysionAI	0.607	0.780	ISIC 2019	✓
Proposed Method	0.593	0.886	ISIC 2019	✗
DermaCode	0.578	0.892	ISIC 2019	✗
Nurithm Labs	0.569	0.870	ISIC 2019 + additional data	✗

5.5 Conclusion

With this chapter we addressed the impact of image resolution, data augmentation, and different state-of-the-art architectures on dermoscopic images analysis. Furthermore, an ensemble strategy that considers augmented samples at test time is presented. Our method successfully deals with the absence of balance between classes, by means of a large use of data augmentation strategies (both at training and testing time), and a weighted cross-entropy loss. The proposed solution takes advantage of multiple networks trained using different augmentation methods. A probabilistic approach is employed to perform the ensemble over calibrated network decisions thus ensuring better results.

Carrying out experiments in a systematic way, we proved that dermoscopic image analysis highly depends on input image resolution, and that the amount of performed data augmentation strategies and available labeled data play a major role in this task. We empirically demonstrated that due to the deficiency of dermoscopic labeled data with respect to natural images, extremely deep architectures (*e.g.* SEResNeXt-101, EfficientNet-b7) fail to provide better results than shallower ones, highlighting that conclusions on natural images can not be directly extended to dermoscopic ones.

The ISIC dataset is expected to consistently keep growing in the approaching years, we therefore plan to extend our studies using larger amounts of data, in order to assess how dermoscopic images analysis is altered if more data is available, if the class imbalance issue is stretched, or if the number of classes is increased. Moreover, state-of-the-art results are obtained by means of extremely expensive inference procedures (both in terms of time and hardware resources). Future research directions will hence include an investigation to find cheaper ways to obtain class predictions from neural networks without excessively lowering the achieved discrimination capabilities. Finally, we plan to explore the effects of new state-of-the-art calibration techniques on an ensemble of neural networks.

The proposed ensemble yielded a balanced accuracy of 0.593 on the official 2019 ISIC challenge, achieving the third best result. It is the best performing method when compared with challenge participants that do not exploit additional data and do not take advantage of an Out of Distribution Data detector. In order to ensure reproducibility, the source-code is provided in [115].

Chapter 6

Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval

6.1 Introduction

A recent international diagnostic study demonstrates that the accuracy of machine learning algorithms for pigmented skin lesion classification proposed in the last years is comparable with those of expert dermatologists [116]. This is true considering only individual images and ignoring the clinical history of the patient: information that the dermatologist has often available when visiting. Nevertheless, in the medical diagnosis field the goal should not be to take an action on behalf of an expert practitioner, but rather to assist his/her choice. Indeed, empirical experiments have shown that providing the physician with a second computerized and non-interactive opinion is not well seen by the dermatologists that even tends to change their own diagnoses when presented with a second computerized point of view [117].

In the last years, many attempts have been made to devise AI ap-



Figure 6.1: Example result of the ranking system. On the left an example query image belonging to the melanocytic nevus class. On the right side the first five results in order of score from top to bottom.

proaches, which are not only performing well, but are trustworthy and transparent for a human expert [118]. Among others, the Grad-CAM technique [65] has proven to be a valid solution for dermoscopic images analysis: it is able to explain dermatologists on which parts of the input image the model is mostly focused, by simply providing a heat map [108].

This chapter mainly aims at providing an interactive approach to effectively support dermatologists in the decision making process. Given a skin lesion, the designed deep learning-based algorithm will retrieve a number of other diagnosed cases that are similar to the original image as shown in Fig. 6.1, thus supporting and driving the dermatologist to a more precise diagnosis, without directly providing a second opinion. We thus propose a Content-Based Image Retrieval (CBIR) tool for searching and retrieving most similar dermoscopic images. A set of quantitative experiments shows that the proposed approach provides a better ranking than existing state-of-the-art solutions on the same topic. Moreover, to test the usefulness of the devised strategy, five dermatologists have been asked to classify different images with and without the output provided by our algorithm: experimental results clearly demonstrate the benefit of the proposal.

The rest of the chapter is organized as follows. The proposed model is described in Section 6.2, and evaluated in Section 6.3. Finally, in Section 6.4 conclusion are drawn.

6.2 Image Retrieval

As stated, the aim of this chapter is to find a model for Content-Based Image Retrieval (CBIR) for skin lesion medical images. Given a query, *i.e.* a new image of an unknown class, the task consists of retrieving k images from a labeled dataset, that are similar to the query. Since the final aim is to provide physicians with a valuable support for the classification task, two images are considered similar if they possess common features ascribable to a certain class of lesions. In order to extract such features, our approach to the problem consists of learning an embedding function f from skin lesion images space I into a compact Euclidean feature space $\mathbb{R}^d$, where distances correspond to a measure of similarity. In feature space, distance between images of the same class should be small, and distance between images of different classes should be large. After such a mapping has been established, the retrieval task can be accomplished by taking the k labeled images with the minimum distance from the query in the embedded space. In order to achieve this task, we built and evaluated four different variations of a deep convolutional neural network.

For every model, input images are resized to 512×512 , and each channel is normalized by subtracting the mean and dividing by the standard deviation. Moreover, training images undergo augmentation techniques to reduce overfitting, consisting in random flip and random rotation.

6.2.1 Classification

The first embedding model is part of a deep neural network trained for classification. We conducted experiments with multiple variations of ResNet (with 18, 34, 50, 101 and 152 layers), which have been proved to be effective solutions for the classification task of the ISIC challenge [77]. The loss used for this network is the cross-entropy loss. The network is trained with Stochastic Gradient Descent (SGD), with an initial learning rate of 0.001, lowered each time the evaluation accuracy reaches a plateau, with initial weights pretrained on the ImageNet dataset. After the training process, the embedding network is obtained by removing the last fully connected layer, which outputs the classification results. In this way, only the feature extraction part of the network is preserved, whose output is a vector of size 512 or 2048, depending on the ResNet variation used. This intermediate feature vector represents high-level visual features, strictly related to the

image class, and represents the mapping in the new euclidean space. This simple model is mainly built for comparison with the similar proposal by Pu *et al.* [57], which also employs a classification network, and with the more complex models discussed further on.

6.2.2 Embedding End-to-End

The second model is a convolutional network trained end-to-end for the embedding task. Like the previous model, it is based on ResNet. However, instead of outputting class probabilities, the last fully connected layer, with output size d , represents the embedding in $\mathbb{R}^d$. The output vector is divided by its L2-norm; this regularization ensures that embedded vectors lie on the surface of the unit $(d - 1)$ -sphere, *i.e.*, $\|f(x)\|_2 = 1$.

The loss used for training is the triplet loss [119], that directly reflects the specific goal of this work. The triplet loss is calculated over a triplet of input samples composed of an anchor (a), a positive (p) and a negative (n). Sample a and p belong to the same class, while n belongs to a different class. Given a triplet, the loss is computed as:

$$\mathcal{L}(a, p, n) = \max(d(f(a), f(p)) - d(f(a), f(n)) + \alpha, 0)$$

In the formula, $d(x, y)$ is the distance between x and y in feature space, and it can be calculated using a chosen metric, while α represents a desired margin between positive and negative pairs. The triplet loss aims at minimizing the distance from anchor to positive, and maximizing the distance from anchor to negative. We performed experiments with the euclidean distance and cosine distance (defined as $1 - \text{cosine similarity}$), and we verified that, for this task, cosine distance yields better results. It is worth noticing that, if feature vectors lie on the unit $(d - 1)$ -sphere, cosine similarity is the same as the dot product.

The training process starts with the sampling of balanced minibatches, with n random elements per class, so that each skin lesion category is equally represented. For every minibatch, a network forward yields a corresponding list of embeddings. Then, triplets are chosen in the following way: (i) for every class, all possible pairs (a, p) , *i.e.* combinations of two samples, are listed. Since there are n samples for each of the 8 classes, the total number of pairs is $\binom{n}{2} \times 8 = 4n(n - 1)$; (ii) for every pair, a hard negative sample is chosen from the other 7 classes. A hard negative is a sample n for which $\mathcal{L}(a, p, n) > 0$.

Only hard negatives are useful for the training process, because they yield a non-zero loss. We experimented the three most common criteria for the choice of the hard negative:

- *random hard*: a hard negative randomly selected.
- *hardest*: the negative sample yielding the highest loss value.
- *random semihard*: a random negative n for which $d(f(a), f(p)) < d(f(a), f(n)) < d(f(a), f(p)) + \alpha$.

The *hardest* negative criterion leads to the fastest convergence, but it is also known to often lead to local minima [120], and the *random semihard* criterion is useful to reduce that risk. These observations hold true in our case, and we got the best results applying the *random semihard* criterion, even if the *random hard* choice proved to be nearly as effective. After the choice of negatives, the loss is computed for every triplet, and losses are averaged to get an aggregated value. This network is trained with SGD and initial learning rate of 0.01, decreased every 30 epochs.

6.2.3 Classification & Embedding

The third model taken into account is composed of both a classification network and a separate embedding network. It builds upon the first model, in the sense that, after the end of the training process, feature vectors extracted by the classification network constitute a new dataset, and they become the input of a fully connected network, that outputs the final embedding in d dimensions. This embedding network is composed of three fully connected layers of size 2048, 1024 and d respectively, randomly initialized. The first two layers are followed by a PReLU activation and a dropout layer; the whole architecture is depicted in Fig. 6.2. Like the previous case, the loss function is the online triplet loss. We trained this network with SGD, learning rate starting at 0.01 and lowered every 30 epochs, and batches of 30 samples per class.

6.2.4 Classification & Embedding End-to-End

The fourth model shares the same architecture of the third one: the difference is that, in this case, the feature extractor and the embedding network are not separated. Instead, after the training of the classification network,

its final fully connected layer is substituted with the embedding network, and a new end-to-end training process starts, with online triplet loss. This means that the backward propagation does not stop at the first layer of the embedding network, as in Model-C, but instead reaches the beginning of the feature extractor. Because of the large size of the network, we are strongly limited in the batch size: in our tests, we used batches with 3 examples for each class. The whole architecture, that is the same as Model-C, is depicted in Fig. 6.2.

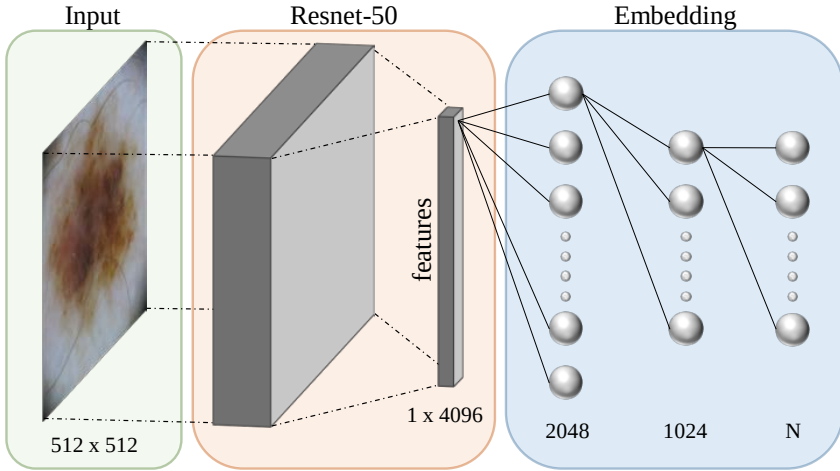


Figure 6.2: Architecture of Model-C (Classification & Embedding) and D (Classification & Embedding End-to-End). The difference between the two approaches is that, after the training of the ResNet feature extractor with a classification loss, Model-D is fine-tuned end-to-end with triplet loss, while Model-C only updates weights on the embedding part of the network.

6.3 Experimental Results

6.3.1 Quantitative

The dataset used for training and evaluating the proposed models is the ISIC 2019 archive [77, 80, 94], an international repository of dermoscopic

Table 6.1: A priori probability of each skin lesion class.

Class	Probability
Melanoma	0.1785
Melanocytic Nevus	0.5083
Basal Cell Carcinoma	0.1312
Actinic Keratosis	0.0342
Benign Keratosis	0.1036
Dermatofibra	0.0094
Vascular Lesion	0.0100
Squamous Cell Carcinoma	0.0248

images built for both clinical training and for supporting technical research by means of the international challenge of the same name, introduced in Section 2.4. The dataset includes 25 331 images, divided into 8 categories: melanoma (MEL), melanocytic nevus (MN), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC) and squamous cell carcinoma (SCC). A priori probability of each class is reported in Table 6.1.

We split the dataset into training, validation and test sets, composed of 19 331, 1 000 and 5 000 images respectively. Every model is trained with the aim of minimizing the corresponding loss function on the training set. After each epoch, a partial evaluation of the model is performed by retrieving the k nearest neighbors taken from the training set of every sample in validation set, for three values of k (1, 5, and 10), and computing precision at k ($P@k$), *i.e.*, the number of neighbors that share the same class of the query, divided by k . The values for k have been chosen in a way that reflects the use case, that is showing a physician similar lesions to a new image, which needs to be classified. $P@k$ values computed on the validation set are used for hyperparameters optimization and early stopping of the training process. Then, final values of $P@k$ are computed by substituting validation set with test set.

In literature, another common metric used to evaluate the performance of retrieval systems is the recall, *i.e.* the ratio of relevant retrieved images to the total number of relevant images in the dataset.

However, the recall is not meaningful for this use case, because the

Table 6.2: Optimal hyperparameters used for each model, found with grid search. FE, Dim, Loss, Lr and BS respectively stand for Feature Extractor, Dimensions of feature space, Loss function, Learning rate and Batch Size.

Model	FE	Dim	Loss	Lr	BS
Classification	ResNet-50	2048	Cross-entropy	0.001	16
Embedding E2E	ResNet-34	300	Triplet	0.001	48
Class & Emb	ResNet-50	30	Triplet	0.01	240
Class & Emb E2E	ResNet-50	30	Triplet	0.01	24

amount of retrieved images is orders of magnitude lower than the total number of dataset images belonging to each class.

As regards the choice of hyperparameters, experiments revealed that, independently of the model, the best value for α is 0.2, cosine distance is more effective than euclidean distance, and the L2 regularization actually improves the results. Model-dependent hyperparameters, instead, are reported in Table 6.2. Feature space dimensionality is fixed for Model-A (*Classification*): it is equal to the size of the second to last layer of ResNet, which precedes the fully connected layer for classification. For the other models, we experimented values 3, 10, 30, 100, 300 and 1000.

It can be observed that the feature extractors used in the final models, ResNet-34 and ResNet-50, include a relatively small number of layers. We verified that deeper versions of ResNet, with 101 or 152 layers, are more effective for the classification task, *i.e.*, lead to higher accuracy; however, when measuring $P@k$, the results are not satisfactory. As regards models employing triplet loss, one of the reasons is that implementation details of the loss itself make large batches more efficient [120], and, under the same hardware, smaller networks allow for larger batches. Another consequence is that Model-C (*Classification & Embedding End-to-End*) reaches lower $AP@k$ than Model-D (*Classification & Embedding*): the latter, despite being a simpler model, can exploit a batch size larger by an order of magnitude.

For comparative reasons, we have implemented the method proposed in [57] and known as Hash-AP. To the best of our knowledge, it is the only skin lesion CBIR system that makes use of deep learning to extract visual features, and it represents the current state-of-the-art in the field. Hash-AP

Table 6.3: Average precision at k measured for every model analyzed, for three values of k , 1, 5, and 10. Central columns report average values separated for each class, and the last column reports the balanced average. Novel proposals are identified by *.

Model	k	AP@ k
Hash-AP [57]	-	0.3404
Hash-AP ResNet*	-	0.7580
Classification*	1	0.8224
	5	0.7859
	10	0.7746
Embedding End-to-End*	1	0.7784
	5	0.7786
	10	0.7804
Class & Embedding*	1	0.8157
	5	0.8168
	10	0.8190
Class & Embedding End-to-End*	1	0.7867
	5	0.7881
	10	0.7920

is a classification network with an added hidden layer that learns a binary embedding, or hash, and retrieves similar images to a query measuring the Hamming distance of hash codes, then clustering the results and the query itself using affinity propagation. The original model is based on AlexNet [14]. In order to provide a fair comparison with our proposals, we also built a variation of Hash-AP based on ResNet, and we verified that, at least on the ISIC dataset, the retrieval performance notably improves w.r.t. the original model.

Final results are reported in Table 6.3 and Table 6.4: $P@k$ is calculated for each class, and the last column reports the average precision at k ($AP@k$), which represents a summary of the model performance. $AP@k$ is the average value of per class $P@k$: in this way, every class is given the same weight, independently of its representativeness. Models Hash-AP

Table 6.4: Average precision at k measured for every model analyzed, for three values of k, 1, 5, and 10. Central columns report average values separated for each class, and the last column reports the balanced average. Novel proposals are identified by *.

Model	k	Per class P@k							
		MEL	NV	BCC	AK	BKL	DF	VASC	SCC
Hash-AP [57]	-	0.4786	0.6111	0.5730	0.1896	0.1984	0.1505	0.3842	0.1375
Hash-AP ResNet*	-	0.8176	0.7558	0.8509	0.7417	0.6256	0.7604	0.8271	0.6851
Classification*	1	0.7840	0.9369	0.9347	0.7400	0.8300	0.7733	0.8667	0.7133
	5	0.7262	0.9111	0.9029	0.7190	0.7724	0.7333	0.8373	0.6853
	10	0.7040	0.9038	0.8957	0.7160	0.7470	0.7213	0.8307	0.6787
Embedding End-to-End*	1	0.7400	0.9018	0.9133	0.6600	0.7520	0.7333	0.8267	0.7000
	5	0.7314	0.8923	0.9005	0.6820	0.7576	0.7387	0.8240	0.7027
	10	0.7322	0.8905	0.8993	0.6855	0.7572	0.7440	0.8253	0.7093
Class & Embedding*	1	0.7490	0.9347	0.8973	0.7150	0.7700	0.8400	0.9067	0.7133
	5	0.7542	0.9347	0.9013	0.7170	0.7768	0.8400	0.9013	0.7093
	10	0.7531	0.9331	0.9032	0.7170	0.7814	0.8453	0.9040	0.7147
Class & Embedding End-to-End*	1	0.7560	0.9022	0.9027	0.6600	0.7600	0.7733	0.8267	0.7133
	5	0.7458	0.9030	0.8992	0.6770	0.7588	0.7707	0.8293	0.7213
	10	0.7436	0.9012	0.9009	0.6830	0.7668	0.7760	0.8373	0.7273

and Hash-AP ResNet retrieve a variable number of images, which depends on the clustering result, and therefore only one value is reported for them. Novel proposals are marked by a star.

The precision measured for the original Hash-AP model is quite low. This result can be attributed to the classification network used, AlexNet, which fails to compete with more modern architectures, and to the difference between the ISIC dataset used in this work and the smaller dataset considered in the original paper. However, the ResNet-50 variation we implemented alongside the original one reaches an average precision of almost 0.76, proving the effectiveness of the idea. As regards models discussed in this chapter, the *Classification* one yields the highest $AP@1$, being the only one to exceed 0.82. Nevertheless, for higher values of k , its performance is less satisfactory and, indeed, this network was not directly trained for image retrieval. The best performing model for higher values of k is Model-C, *Classification & Embedding*. As stated before, this result is partially related to the batch size used in training: Model-C has been trained with a batch size respectively 5 times and 10 times larger than that used for Model-B and D (Table 6.2). A larger batch size means that, in the triplet selection process, more negatives are evaluated for the same positive pair, and thus there is a higher probability to find a *random semihard* negative. The improvement with respect to Hash-AP ResNet demonstrates the effectiveness of the triplet loss function, that is more suitable than a classification loss for the retrieval task.

6.3.2 Qualitative

In order to understand if our model can be effectively useful in supporting medical diagnosis, we asked five dermatologists to undergo a test, composed of two tasks. The first task simply consisted in classifying 100 dermoscopic images, randomly sampled from the ISIC test set, without any additional help. In the second task, the physicians were asked to classify the same images, but we also showed them the 5 nearest neighbors of each sample, labeled with their ground truth classes. The neighbors were found in the training set using Model-C. The tests were performed using exactly the same 100 images, with the aim of reducing the probability of a possible improvement being only imputable to the dataset variance. In order to mitigate bias in the second task, each image was rotated by 180°, and the list was shuffled. We measured the accuracy of the physician classification

Table 6.5: Accuracy of 5 dermatologists on Task 1 and Task 2, consisting in classifying 100 skin lesion images with and without the support of our retrieval network.

	Task 1	Task 2
Dermatologist #1	75%	79%
Dermatologist #2	64%	80%
Dermatologist #3	69%	71%
Dermatologist #4	68%	82%
Dermatologist #5	61%	71%

for both tasks. Average result is 67.4% for the first one (images only) and 76.6% for the second one (images and 5 nearest neighbors). Results obtained by individual dermatologists are shown in Table 6.5. Although a set of five dermatologists constitutes a relatively small statistical sample, an average improvement of 9.2% on the very same images proves the usefulness of our CBIR model as a support in the decision making process.

6.4 Conclusion

In this chapter we introduced a novel content-based image retrieval system for dermoscopic image analysis, with the aim of addressing the low-interpretability issue that burdens deep learning methods and build a tool that could be useful for human dermatologists. The proposed system is based on learned features obtained through state-of-the-art convolutional neural networks.

An exhaustive experimental evaluation proved that the proposed solution outperforms competitors on the well known ISIC dataset. Moreover, the effectiveness and the usefulness of the devised solution have been demonstrated by a set of qualitative experiments carried out by a group of dermatologists.

These results prove that, although the existing knowledge of dermatologists is invaluable for the diagnosis of skin cancer, the proposed CBIR model can assist physicians with dermatologist-grade decision support.

Chapter 7

Where Do Neural Networks Look?

7.1 Introduction

The great discriminative power of neural networks comes at the cost of very low explainability, hence it is extremely difficult to understand the reasoning behind a model prediction [105]. Therefore, the possibility of CNNs learning a bias within the dataset instead of meaningful patterns has been thoroughly investigated by several authors [9, 61]. In particular, we consider “bias” any dataset-intrinsic characteristic that can be exploited by a neural network to learn the underlying distribution, but that is not related to the class and thus greatly hinders the generalization capabilities of the algorithm [121]. In other words, our goal is to explore the problem of shortcut learning in dermoscopic images analysis [122]. This thesis focuses on the dermoscopic images domain where, fortunately, relevant characteristics to be considered when analyzing skin lesions have been explicitly outlined by expert practitioners. As a matter of fact, five main components for melanoma detection have been identified –and are used daily by dermatologists– in the ABCDE rule [47]. These characteristics are *Asymmetry*, *Border* irregularity, *Color* variegation, *Diameter* (> 6 mm) and *Evolution*, and it is desirable for automatic classification algorithms to focus on them instead of considering irrelevant artifacts such as checker-

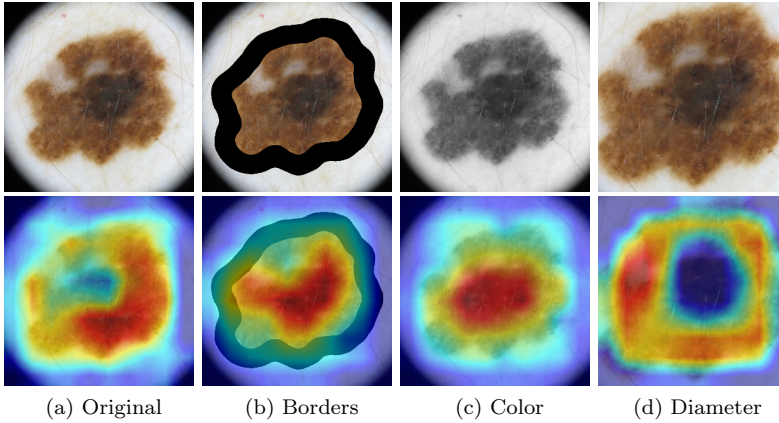


Figure 7.1: Grad-CAM visualization when debasing different ABCD properties. More examples are provided in Fig.7.3.

board patterns introduced by sharpening filters, black round borders, pen drawings, rulers, and hair, which should be ignored when seeking meaningful elements within an image. This chapter explores the correlation between CNNs performance and traditional dermoscopic features, by carefully evaluating the impact of debasing semantically important components from images. Indeed, each ABCDE component is individually “removed” in order to study the discriminative power of clinical characteristics in this controlled environment, where very important pieces of information are denied to classification algorithms (Fig. 7.1).

Most importantly, the experiments performed in this chapter are built on two different datasets, and thus yield very valuable insights about the biases potentially present within a single set of data.

The rest of this chapter is organized as follows. Section 7.2 describes the datasets employed for carrying out the proposed studies, which are motivated and described in detail in Section 7.3. Experiments and their methodology are reported in Section 7.4, while Section 7.5 presents an exhaustive analysis of the results and, finally, conclusions are drawn in Section 7.6.

7.2 Dermoscopic Images

The International Skin Imaging Collaboration (ISIC) began to aggregate a large-scale, publicly available collection of dermoscopic skin lesion images in 2016 [77]. This *easy-to-reach* considerable amount of dermoscopic images allowed to significantly improve the performance of machine learning algorithms. The dataset –also known as the ISIC archive– is designed for clinical training and, mostly, to support research toward automated skin cancer analysis, as showcased by several challenges and workshops hosted over the years. The 2019 version of the ISIC archive contains a total amount of 25 331 labeled dermoscopy images belonging to nine different classes [94], which represent eight types of skin lesion plus an additional category, not available in the training partition of the data. This last class, named *none of the others*, contains dermoscopic images of different natures that do not belong to any of the other eight classes. Images have been collected over a span of many years, from multiple centers, and using different devices.

Table 7.1: Class distribution of the three employed datasets shown as percentages: ISIC 2019 and 2020, and the private dataset.

Class	Label	2019	2020	Private
Melanoma	MEL	17.8	1.8	16.7
Melanocytic Nevus	NV	50.8	15.7	58.1
Basal Cell Carcinoma	BCC	13.0	–	7.6
Actinic Keratosis	AK	3.0	–	1.6
Benign Keratosis	BKL	10.0	0.7	6.3
Dermatofibroma	DF	0.9	–	1.0
Vascular Lesion	VASC	1.0	–	0.0
Squamous Cell Carcinoma	SCC	2.4	–	1.8
Unknown	unknown	–	81.9	6.9
Total		25 331	33 126	25 849

As depicted in Table 7.1, the available data is heavily imbalanced. Therefore, the 2019 official metric was the balanced accuracy metric, which is computed as the average sensitivity among classes regardless of how much each class is represented in the test set. Additional metrics such

as the Area Under the ROC Curve (AUC) are presented in the official leaderboard, but not employed for the final scoring.

On the other hand, the 2020 challenge edition, hosted on Kaggle¹, switched to a binary classification problem: melanoma or non-melanoma. Thus, the evaluated metric has been changed accordingly, leading to the adoption of AUC. Moreover, the 2020 SIIM-ISIC challenge dataset [123] gained patient-level contextual information, providing for each image an identifier which allows lesions from the same patient to be mapped to one another. This additional knowledge is frequently used by clinicians to diagnose melanoma and is especially useful in ruling out false positives in patients with many atypical nevi. This new dataset is composed of 33 126 images and collected from 2 056 patients gathered from 1998 to 2020 (21% of them with at least one melanoma, 79% with zero melanomas), with an average of 16 lesions per patient. Even though the challenge focused on benign/malignant lesion classification, the data reports the exact lesion diagnoses, such as melanoma, nevi, atypical melanocytic proliferation, café-au-lait macule, lentigo NOS, solar lentigo, lichenoid keratosis, seborrheic keratosis, and the aforementioned “unknown” class.

7.2.1 Private Dataset

In order to determine the existence of biases in the public ISIC dataset, we extend the experiments by means of a private dermoscopic dataset (Fig. 7.2) consisting of 25 849 images, collected between 2003 and 2019 using several distinct acquisition tools. Sample classes are mapped into the ISIC2019 diagnostic categories, and images that do not belong to any of the other eight classes are assigned the “unknown” label. As shown in Table 7.1, the dataset presents a different category distribution compared to the ISIC2020 dataset, with a higher percentage of melanoma cases. Similar to both ISIC datasets, the private collection of data contains several clinical information such as sex, age and site of the lesion. Contrarily to the public ISIC dataset, visual artifacts which could be considered source of shortcut learning, such as rulers, ink markings/staining, and colored patches are almost completely absent in our private dataset. The whole set of dermoscopic images is used as an additional test set for the experiments and analyses carried out in this chapter, and thus yields important

¹kaggle.com/c/siim-isic-melanoma-classification

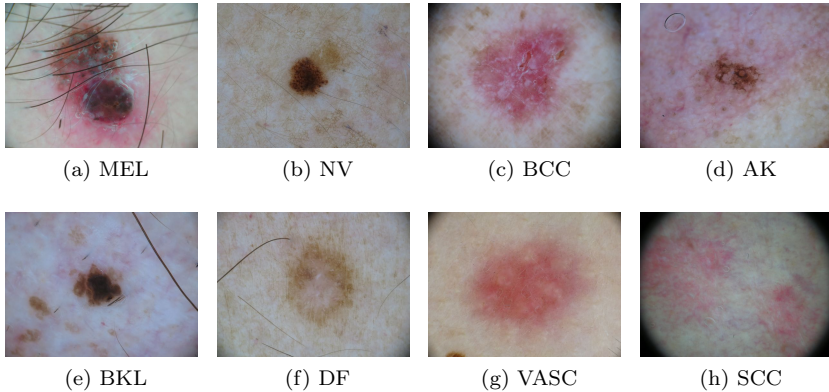


Figure 7.2: Samples of the private dataset. (a) Melanoma - MEL, (b) Melanocytic Nevus - NV, (c) Basal Cell Carcinoma - BCC, (d) Actinic Keratosis - AK, (e) Benign Keratosis - BKL, (f) Dermatofibroma - DF, (g) Vascular Lesion - VASC, (h) Squamous Cell Carcinoma - SCC.

information about possible irrelevant and for-class-consistent characteristics within the ISIC dataset, which might be learned and exploited by CNNs.

7.3 Investigating ABCDE features

The aim of the proposed chapter is to investigate the presence of biases in dermoscopic images datasets. A *dataset bias* may be defined as a collection of features that are semantically irrelevant for the investigated task, but which can be exploited by neural networks to improve the evaluation metrics. To name a few, CNNs might pick up common characteristics within malignant skin lesion images, such as artifacts related to pen drawings, markers, colored patches, or rulers. Moreover, deep learning algorithms are able to recognize acquisition device models and calibration settings, thus identifying the provenience of an image which might be highly related to the final diagnosis [59]. On the other hand, neural networks are expected to focus on relevant features for dermoscopic image analysis, which can be summarized as *Asymmetry*, *Border* irregularity, *Color* variegation, *Dia-*

meter, and *Evolution*: the ABCDE rule [124, 47]. Hence, it is extremely important to be able to interpret which image characteristics neural networks take into account when making a class prediction. This can be achieved by means of Class Activation Mapping (CAM) strategies. In particular, experiments in this chapter employ Grad-CAM [65] to locate the regions of an image that most contribute to the final prediction. Moreover, we run an extensive set of experiments to study how introducing noise in the ABCDE properties affects neural network performance, and analyze which sections of an image CNNs focus on when crucial features are debased or removed. Some of the experiments described in the following exploit segmentation masks obtained by means of DeepLabv3+ [30], trained using the 2017 ISIC segmentation task dataset [108]. Samples images obtained through feature debasing are reported in Fig. 7.3 and generated by means of the European Computer Vision Library –ECVL– [113].

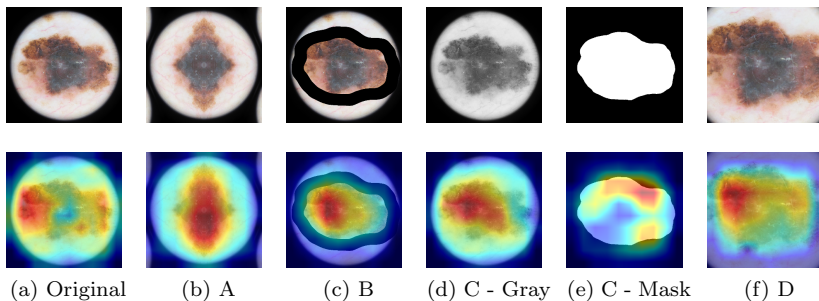


Figure 7.3: Grad-CAM visualization when debasing different ABCD properties.

7.3.1 Tampering with Asymmetry

Asymmetry is one of the most important visual features for melanoma detection [19], it can be described as the difference in volume and shape of two parts of a skin lesion, obtained by *cutting* it with a straight line passing through its center. In order to train a symmetry agnostic neural network, dermoscopic images can be split by a random straight line and by its perpendicular, both passing through the center of the lesion. Subsequently,

a quarter of the image can be flipped over both axes in order to obtain a version of the original lesion with increased symmetry. Practically, the center of the image is aligned with the centroid of the lesion, obtained from the segmentation mask. The image is then randomly rotated and, finally, the top-right quarter is flipped with respect to the horizontal and vertical axes.

7.3.2 Concealing Borders

With the aim of removing valuable information about the shape of a skin lesion edge, which is a crucial aspect when assessing its malignancy, we cover borders with a thick black line, obtained from the contour of the segmentation map. Firstly, a morphological dilation operation is applied to the contour, with a kernel size proportional to both the image size and the foreground-background ratio. Then, with the goal of smoothing out irregular segments, a Gaussian filter with a large kernel is applied. Finally, the black border image is superimposed on the original one, thus removing any information about the actual transition from the human skin (background) to the actual lesion (foreground).

7.3.3 Removing Color

The presence of multiple colors within a single mole (blue, black, white, red, and brown) or the uneven distribution of color can sometimes be a warning sign of melanoma, since most benign lesions are usually a single shade of brown or tan [125]. Two different sets of experiments are conducted in order to assess the effects of discarding information about color from dermoscopic images. The first one is run by simply converting the image from RGB to grayscale, thus removing any knowledge about the different colors within a skin lesion. However, while this processing step erases any data about hue and saturation, it does not affect the luminance, thus leaving the CNN the chance to learn valuable features from the color distribution within moles. In the second experiment, color features are completely removed, as we train a neural network to classify skin lesions using only their segmentation masks. In this extreme setup, the neural network is fed with minimal knowledge about skin lesions, and is forced to make a prediction based uniquely on noisy, automatically obtained, binary mole shapes.

7.3.4 Altering Diameter

Because skin cancer cells grow abnormally fast, diameter is one of the most important parameters in skin lesion classification. Unfortunately, dermoscopic images are acquired at several scales, which are not always included as metadata and can not be deducted from the image, as only a limited amount of samples contain a ruler. As a matter of fact, a mole that exceeds the borders of the image is not necessarily larger than one that does not; information about diameter is extremely noisy and very hard to investigate, yet potentially extremely important for melanoma detection algorithms. To remove any information about mole dimensions, the foreground-background ratio of images is set to a fixed percentage. We choose to train a CNN uniquely with samples where the skin lesion represents the 80% of the image, as it yields good qualitative results. In order to achieve this, samples where the skin lesion is contained in just a small portion of the image must be cropped, using the mole centroid as the center. On the other hand, images where the mole fills more than 80% of the original sample are padded, by reflecting the sections of the image closest to the borders. As shown by the foreground density histograms in Fig. 7.4, this method mostly results in crops, whereas only a very small portion of the dataset (foreground percentage $> 80\%$) needs padding. By following the aforementioned *Crop&Pad* technique, in the rare cases of very elongated lesions, a small part of the mole is left out by the crop. However, each image will present the same number of foreground pixels, thus eliminating the image scale differences and bringing all the lesions to the same size.

7.3.5 About Evolving

Dermoscopic images seldom contain information about evolution, as only in few cases follow-up data is provided in the existing datasets. Introducing such additional data in dermoscopic dataset would certainly have significant implications in the research field, but it cannot be considered and analyzed nowadays.

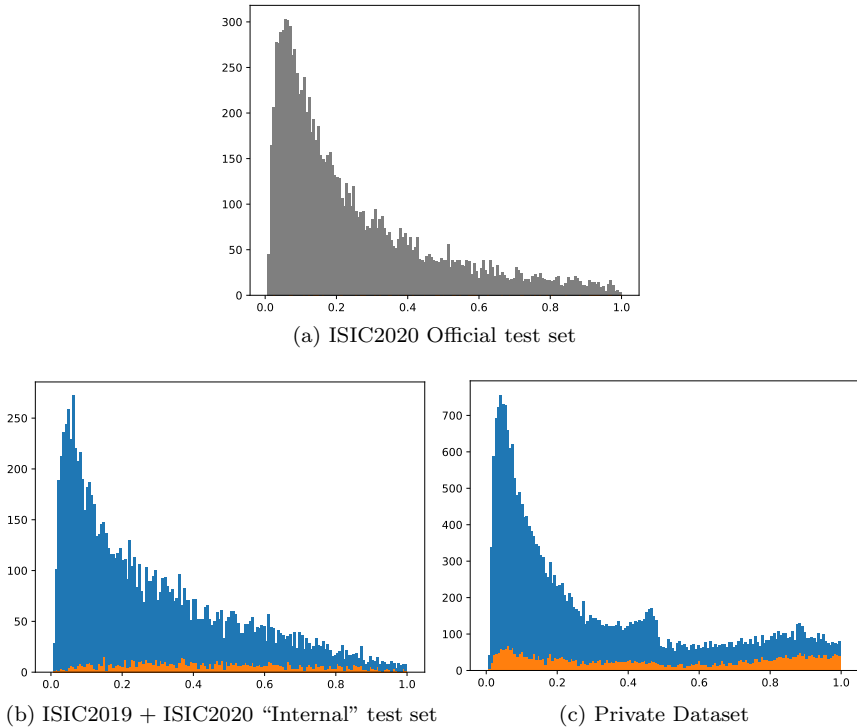


Figure 7.4: Histograms of foreground density distribution within different test sets. Benign and malignant skin lesions are depicted in blue and orange respectively. Best viewed in color.

7.4 Experiments

To begin with, an ensemble of neural networks has been developed in order to classify melanoma within the ISIC2020 challenge hosted on Kaggle. Our algorithm is inspired by the winner of the challenge [126], and it obtains basically the same results. The dataset employed for all the experiments combines ISIC2019 and ISIC2020 data, cleaned from duplicated [127] for a total of 57 964 training images. The 2020 classes are mapped into 2019 ones.

In order to compensate for dissimilarities between image sizes, a squared

center crop is performed to produce images of $\min(h, w) \times \min(h, w)$ pixels, later resized to 768×768 .

Metadata such as sex, age, and site of the lesion, are provided for both datasets. In ISIC2020, each image is also associated to the number of patient lesions,

since it could help machine learning algorithms to identify a pattern known in dermatology as the “ugly duckling sign” [128], which is frequently used by physicians to diagnose melanomas. In the ISIC2019 dataset, for which the *patient id* is not provided, we set the same metadata to 1.

The ISIC2020 challenge winning method needs to train 18 neural networks for 15 epochs to obtain the best AUC result of 0.9490. With the aim of obtaining comparable results, but reducing the computational load in terms of time and memory, only 6 networks have been selected based on the best trade-off between training time and performance, pointed out in [129]. Each of them is powered by a different setup, shown in Table 7.2. Images can be pre-processed with an *essential* augmentation pipeline that includes only transpose, flip, or with a *full* one which combines transpose, flip, rotate, random brightness, random contrast, additive motion blur, median blur, gaussian blur, gaussian noise, optical distortion, grid distortion, elastic transform, CLAHE, hue-saturation-value modification, random affine transformations, and dropout of square regions. Both pipelines eventually resize and normalize the input images.

Each network is trained with the Cross-Entropy loss and Adam optimizer [76], with a learning rate of 3×10^{-5} . Metadata are fed to two fully connected layers whose output is concatenated to the one of the convolutional network, and then passed to a final fully connected layer, which performs the classification. Each network is trained for 20 epochs and produces 9 class probabilities as output, among which only the melanoma class is considered. In order to take advantage of all the training data and obtain more reliable validation results, a 5-fold cross validation procedure is exploited by taking about 1/5 of the images as validation, maintaining for each fold the same percentage of melanomas as in the training set (Table 7.1).

The final ensemble probabilities are obtained by averaging the predictions of each model, weighted by their rank in $[0, 1]$. This allows to mitigate problems related to different model calibration.

The aforementioned process allowed us to obtain an AUC score in the private challenge leaderboard of 0.9469, which is comparable with the

state-of-the-art of 0.9490, obtained with an ensemble of 18 neural networks instead of just 6. Experimental results obtained by single network configurations are reported in Table 7.2.

As already discussed, the goal of this chapter is to investigate any possibility of shortcut learning within the ISIC dataset, by studying the correlation between CNNs performance and traditional dermoscopic features. In order to produce a comprehensive evaluation, multiple sets of experiments are required, making it unfeasible to train the entire ensemble for each of them. For this reason, the computation load has been reduced by selecting the lightest setup involved in our ensemble: EfficientNet-B3 with image size of 300×300 , without 5-fold cross-validation.

The ISIC2020 test set labels are private, and the Kaggle challenge platform only evaluates algorithms with the AUC metric. Thus, in order to extend our analysis by including additional evaluation metrics, the validation set has been split to produce an “internal” test set: the resulting dataset counts 46 379 training images, 1 159 for validation and 10 426 images for testing. The *Private Dataset* is employed for testing classification performance as well. All of these images have been processed using in turn the techniques described in Section 7.3 to remove symmetry, lesion border, color (using grayscale conversion or directly the segmentation mask) and to change the lesion diameter. The results are five different datasets, used to train and test the selected architecture with the aforementioned setup. Experimental results obtained on the ISIC2020 “internal” test set and on the private dataset are reported in Table 7.3.

For the sake of completeness, Table 7.4 reports the results obtained on the official ISIC2020 test set, evaluated with the AUC metric score provided by Kaggle.

Table 7.2: Description and setup of the models employed for the final ensemble. The experimental results obtained by using a single neural network on the ISIC2020 official test set and Private dataset are respectively reported in the last two columns.

Net	Image size	Batch size	Metadata	Augmentations	AUC ISIC	AUC Private
EfficientNet-B3	512×512	20	✓	Full	0.9340	0.8895
EfficientNet-B3	300×300	60	✗	Full	0.9049	0.7983
EfficientNet-B4	448×448	20	✗	Full	0.9338	0.8823
EfficientNet-B5	384×384	20	✗	Full	0.9317	0.8804
EfficientNet-B5	768×768	8	✓	Full	0.9337	0.8982
ResNet-152	256×256	40	✗	Essentials	0.9281	0.8440
Ensemble					0.9469	0.8703

Table 7.3: Experimental results using the different input configurations described in Section 7.3 on the ISIC2019 + ISIC2020 “internal” test set, and on the private dataset.

Experiment	AUC	Precision	Recall Sensitivity	Specificity	FScore	Accuracy
ISIC2019 + ISIC2020 “internal” Test Set						
Original	0.9671	0.7821	0.7180	0.9808	0.7487	0.9577
Asymmetry	0.9448	0.7755	0.5399	0.9850	0.6366	0.9459
Borders	0.9605	0.7326	0.6678	0.9766	0.6987	0.9495
Color (Grayscale)	0.9559	0.7420	0.7071	0.9763	0.7241	0.9527
Color (Mask)	0.8017	0.6897	0.0656	0.9972	0.1198	0.9154
Diameter	0.9724	0.8216	0.7399	0.9845	0.7786	0.9631
Private Dataset						
Original	0.7983	0.5299	0.5038	0.9104	0.5165	0.8425
Asymmetry	0.7693	0.5553	0.4025	0.9354	0.4667	0.8465
Borders	0.7896	0.5261	0.4992	0.9099	0.5123	0.8413
Color (Grayscale)	0.7673	0.4607	0.4540	0.8935	0.4573	0.8201
Color (Mask)	0.7032	0.6017	0.0322	0.9957	0.0612	0.8349
Diameter	0.8099	0.5597	0.5168	0.9185	0.5374	0.8515

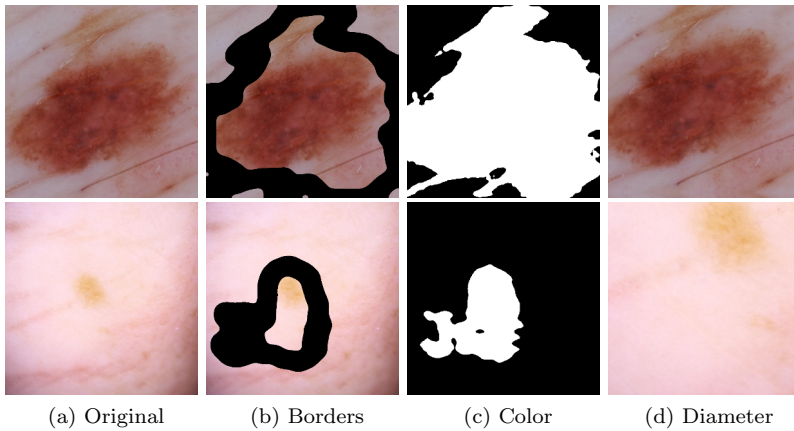


Figure 7.5: Example of failure cases, due to wrongly generated masks.

7.5 Discussion

Almost every experiment described in Section 7.3 is built on automatically drawn segmentation masks, which inevitably introduces sporadic failure cases. As a matter of fact, when segmentation masks are not correct our image processing techniques fail to hide the considered feature (Fig. 7.5). However, the proposed ABCD-concealing processing methods opportunely obscure important information about ABCD features in most scenarios, and thus successfully force neural networks to not focus on the degraded characteristic (Fig. 7.3). In particular, in Fig. 7.3c the neural network trained to classify images with hidden borders makes a prediction focusing on the section of the lesion with most variance of color intensity. The same patch is equally important for the network fed with grayscale images (Fig. 7.3d), whereas Fig. 7.3a clearly shows that mole borders are of great interest for the “standard” model. On the other hand, the CNN asked to make a prediction based solely on the segmentation mask strictly focuses on the sections of the foreground with higher concavity, which is roughly the only *valuable* piece of information about the lesion to be found in the extremely degraded input. As suggested by the high classification performance, neural networks autonomously learn to extract features that

Table 7.4: Experimental results using the different input configurations described in Section 7.3 on the ISIC2020 official test set.

Experiment	AUC ROC
Original	0.9049
Asymmetry	0.8921
Borders	0.8752
Color (Grayscale)	0.8753
Color (Mask)	0.7831
Diameter	0.8818
Mask Area	0.7421
Bounding Box Area	0.7375
Bounding Box 70% Area	0.6599

are notoriously important for melanoma detection.

The good discriminative power of CNNs is confirmed on the private dataset, despite the slight degradation *wrt* ISIC test sets. It is worth noticing that CNN performance tend to naturally drop when source domain (training data) and target domain (test data) come from distinct origins, even for extremely simple tasks such as handwritten digit recognition, where classification accuracy across separate datasets can be decreased by up to 40% [130]. This performance drop can be due to a large number of reasons, such as different lighting settings, resolution, image quality, human-introduced artifacts, subject centering, and image acquisition devices [131, 132].

7.5.1 A Crucial Characteristic in Dermoscopic Datasets

However, our observations clash with the conclusion drawn by Bissoto *et al.*[61], which shows that CNNs achieve good performance even when skin lesions are completely covered with black patches. The authors also bring their experiment one step further by fixing the dimension of the black patches as $\geq 70\%$ of the image, thus covering the vast majority of input samples, and still reach a surprisingly high AUC of 0.712. We make the assumption that foreground-background ratio plays a crucial role in this

round of experiments, since the size of the bounding boxes around skin lesions remains available to trained CNNs. Hence, rather than finding “a pattern within the available pixels [not hidden by the black patch] that contain information that leads to the correct label” as assumed by Bissoto *et al.*, neural networks can simply make a prediction based on the mole size. Actually, foreground densities obtained from segmentation masks can be used as malignancy prediction. The *Segmentation Mask* experiments reported in Table 7.5 show that the AUC obtained using only the foreground density as malignancy prediction is comparable with that of neural networks trained on skin lesion samples patched with black boxes [61]. This experiment has been pushed further by making predictions based only on lesion bounding box dimensions and, finally, by setting the foreground-background ratios as $\geq 70\%$. Results obtained are reported in the aforementioned table with the name of *Bounding Box* and *Bounding Box 70%*.

These clearly demonstrate that the foreground size within an image is sufficient to obtain a “decent” AUC in the melanoma detection task, even when degraded by substituting segmentation masks with bounding boxes or when discarding any information about images where the lesion bounding box occupies less than 70% of the entire image.

Finally, histograms in Fig. 7.4 show beyond any doubt that the probability of a lesion being malignant grows with the dimensions of the lesion within a dermoscopic image. Intuitively, a mole that exceeds the borders or gets very close to them is not necessarily larger than others, but it is more likely to be malignant. This characteristic might be more related to the complexity of including a whole malignant lesion when acquiring dermoscopic images [133, 134], than to the diameter itself. Nevertheless, this feature is strongly related to the nature of dermoscopic images, and is present across different datasets.

7.5.2 About AUC

The Area Under the Receiver Operating Characteristic curve (AUC) is a well-known metric designed to evaluate the diagnostic capabilities of binary classifiers. It is the official metric of the ISIC2020 challenge, and it offers the advantage of not needing a fixed threshold, thus supplying one less parameter to “overfit” proposed algorithms on the official test set with. However, real clinical applications require a threshold to be set, and a class prediction to be given; evaluating experimental results uniquely by

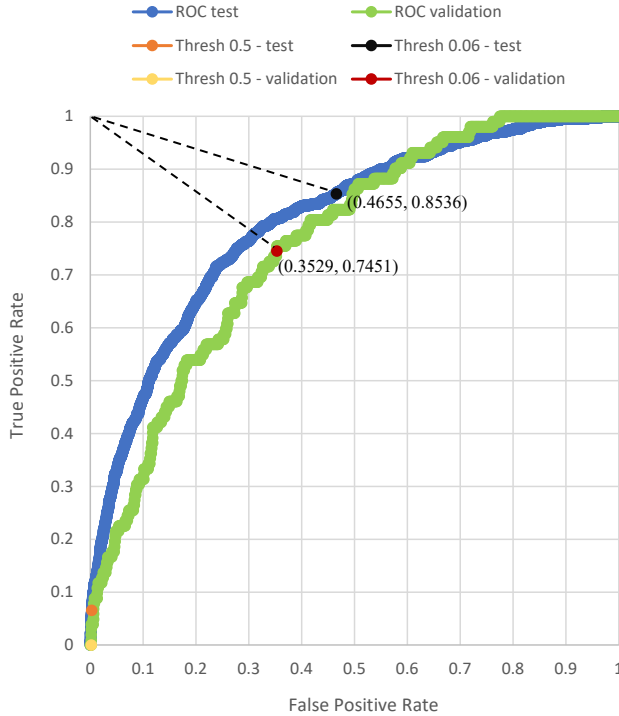


Figure 7.6: ROC curves for the *Mask* experiment on the ISIC2019 + ISIC2020 “internal” validation and test sets (Table 7.3 and Table 7.6). The threshold value that minimizes the distance from the (0;1) point in the validation set is exploited for the test set, corresponding points are highlighted in both ROC curves. Moreover, the points corresponding to the 0.5 threshold are highlighted in the two curves.

means of the AUC metric can be misleading. To put results into context, we further discuss the performance of the CNN trained to classify skin lesion binary masks. Table 7.3 shows that the investigated network yields an AUC of 0.8017 when tested on the subset of public images (ISIC2019 and ISIC2020) used as “internal” test set. This is the Area under the blue curve depicted in Fig. 7.6. When following this strategy, we obtain a

tool with sensitivity 0.0656, and specificity 0.9972, which means that the model is able to correctly recognize only 6.5% of the melanoma cases, but successfully identifies 99.7% of the not-melanoma cases. Clearly this is not the positive result that an AUC of 0.8017 might suggest.

As a matter of fact, the assumption that 0.5 is an appropriate threshold when dealing with neural networks is not correct [60], as showed by Fig. 7.6. Alternatively, the threshold can be set by studying the ROC curve obtained on the validation set (green curve in Fig. 7.6), and choosing the value in the graph closer to point $(0; 1)$, *i.e.* the value that maximizes *True Positive Rate* while minimizing *False Positive Rate*. In this particular case, the desired rate is ≈ 0.06 , and by employing this same threshold on the CNN outputs over the test set we obtain a binary classifier with sensitivity 0.8536 and specificity 0.5345. Table 7.5 and Table 7.6 present the results obtained by setting the prediction threshold following the described steps, always making use of the validation set. Regardless of how thresholds are set, it is clear that high AUC values do not always correspond to satisfying discriminative capabilities.

Table 7.5: Experimental results using foreground densities obtained from segmentation masks, bounding boxes and bounding boxes that cover at least 70% of the image as melanoma probability on the ISIC “internal” test set and on the private dataset.

Experiment	AUC	ROC	Precision	Recall Sensitivity	Specificity	FScore	Accuracy
ISIC2019 + ISIC2020 “Internal” Test Set							
Segmentation Mask	0.7215		0.1483	0.7388	0.5917	0.2470	0.6046
Bounding Box	0.7154		0.1483	0.7202	0.6019	0.2459	0.6123
Bounding Box 70%	0.6220		0.1830	0.3989	0.8286	0.2509	0.7909
Private Dataset							
Segmentation Mask	0.6980		0.2856	0.5898	0.7043	0.3848	0.6852
Bounding Box	0.6919		0.2573	0.6589	0.6190	0.3701	0.6256
Bounding Box 70%	0.6517		0.3328	0.4735	0.8098	0.3909	0.7536

Table 7.6: Experimental results using the different input configurations described in Section 7.3 on the ISIC “internal” test set and on the private dataset. Each experiment uses its own threshold, calculated as the value of the ROC curve which minimizes the distance from (0;1).

Experiment	AUC	ROC	Precision	Recall Sensitivity	Specificity	FScore	Accuracy
ISIC2019 + ISIC2020 “internal” Test Set							
Original	0.9671		0.3163	0.9519	0.8020	0.4748	0.8152
Asymmetry	0.9448		0.2527	0.9628	0.7260	0.4003	0.7468
Borders	0.9605		0.2218	0.9858	0.6672	0.3621	0.6952
Color (Grayscale)	0.9559		0.2590	0.9628	0.7349	0.4082	0.7549
Color (Mask)	0.8017		0.1500	0.8536	0.5345	0.2551	0.5625
Diameter	0.9724		0.2419	0.9803	0.7044	0.3881	0.7287
Private Dataset							
Original	0.7983		0.2578	0.8570	0.5055	0.3963	0.5642
Asymmetry	0.7693		0.2140	0.9017	0.3365	0.3460	0.4308
Borders	0.7896		0.2012	0.9300	0.2600	0.3308	0.3718
Color (Grayscale)	0.7673		0.2291	0.8844	0.4038	0.3639	0.4840
Color (Mask)	0.7032		0.2421	0.7604	0.5229	0.3672	0.5626
Diameter	0.8099		0.2150	0.9263	0.3221	0.3489	0.4230

7.6 Conclusion

With this chapter, we explored the correlation between automatic skin lesion classification and the ABCDE rule. This was done by gradually removing important information from CNN inputs, and by means of state-of-the-art explainability tools such as Grad-CAM [65]. Experimental results show that neural networks autonomously learn to extract features that are notoriously important for melanoma detection, but also prove that their performance are still satisfying when some of these features are removed. This should by no means be considered shortcut learning: instead, remaining information is simply enough to achieve satisfying or even good classification accuracy.

Moreover, we test the model performance on a totally distinct private dataset, with no possible intersection with samples employed during the training phase, showing that state-of-the-art algorithms are not affected by intra-dataset biases. Indeed, this domain is characterized by “simple” attributes with meaningful discriminative information. In particular, the proposed chapter experimentally proved that the foreground-background ratio is strongly related to the malignancy probability of a skin lesion. The reasoning behind this might be related to the well-known *diameter* characteristic from the ABCDE rule, but also to the fact that capturing the entire malignant mole in a dermoscopic image is usually not trivial given its dimensions, the non-clearly defined borders, and the irregular shapes that characterize cancerous skin lesions [133, 135, 134, 136]. Nevertheless, foreground-background ratio is a valuable dermoscopic property.

Finally, the employed explainability tools showed that state-of-the-art architectures focus on the actual lesion and, when important information is concealed, they learn to detect melanomas through the remaining features.

Chapter 8

Conclusion

The goal of the research activities carried out during my Ph.D. was to develop deep learning techniques for skin lesion analysis and automated diagnosis systems. That goal has been tackled by following three different themes which guided all the work presented in this thesis.

First of all we focused on the segmentation task, a fundamental step in the automated melanoma detection process. We approached this task by tackling the huge need of medical manually annotated data by means of a GAN-based novel data-augmentation technique. This algorithm can be used to pre-train segmentation neural networks, and it proved to be particularly effective in data-scarce environments. Moreover, this pre-training approach was employed in a diverse ensemble technique to raise the state of the art in skin lesion segmentation.

We then focused on the skin lesion classification task. Carrying out experiments in a systematic way, we proved that input image resolution is by far the most important scaling-up hyperparameter in automated skin lesion diagnosis. We empirically demonstrated that conclusions on natural images can not be always extended directly to dermoscopic ones. By following the guidelines set in this work, the third best result in the official 2019 ISIC challenge was achieved.

Finally, this thesis addressed the low-interpretability issue that hinders deep learning algorithms. In order to do so, we proposed a content-based image retrieval system to provide a transparent aiding tool, which can assist physicians with dermatologist-grade decision support. Moreover, we

explored the relation between automated diagnosis systems and the well-known dermoscopic features on which human dermatologists base their examination. A very strong connection was detected between the foreground-background ratio of a skin lesion image and its probability to be malignant.

Despite all of the efforts spent in recent years, CNN interpretability remains a massive open question for dermoscopic image analysis as much as for countless additional computer vision tasks. As novel techniques for interpretable neural networks are developed, it is crucial to include them in dermoscopic image analysis and exploit them to develop valuable, easy-to-use tools for human practitioners.

An additional interesting future direction for computer vision in dermoscopic image analysis would be building tools for the training of young dermatologists, and one way to do it would be through the most recent generative models. Photo-realistic synthetic images can be generated with specific characteristics, and thus employed to train humans in identifying malignant lesions with an increasing level of difficulty. Additionally, novel generative models could be designed to simulate the evolution of existing skin lesions, hence creating helpful data to be used in the humans training process.

Finally, an important future direction to explore for deep learning in dermoscopic image analysis would be learning to score skin lesions not on their own, but with the aid of additional lesions present on the same patient. This is a meaningful step for human practitioners, which on a daily basis use similarity between multiple lesions on the same patient as a crucial feature for diagnosis.

Appendix A

Additional Experimental Results

Some of the experimental results discussed in the previous Chapters have been placed in this Appendix to make the reading flow of the thesis easier. They are here reported divided by Chapters.

Augmenting data with GANs to segment melanoma skin lesions, 3

A large dataset is a crucial asset for any GAN training process: additional images allow the network to learn how to generate more realistic samples, with realistic details and less similar one another. Fig. A.1, Fig. A.2, and Fig. A.3 show how enlarging the training dataset improves the output results of the DCGAN. Increasing the ammount of samples from 500 to 1000, the DCGAN becomes able to provide a wider variety of skin lesions, with different sizes and shapes and with various textures. After reaching 1500 training images, the framework delivers high-resolution hair in samples that present much fewer noisy artifacts.

The sampling process of the LAPGAN is further examined in Fig. A.4. We merge five independent GANs to form one LAPGAN, which is divided into six different pyramid levels. In the first level, the GAN named G_0 transforms a noise vector Z_0 into a 6×8 pixels skin lesion sample coupled with its segmentation mask, employing fully-connected layers for both the generator and the discriminator.

In the next pyramid level, the two outputs of G_0 are upsampled and fed, together with a new source of noise, to G_1 , a GAN that exploits convolutional layers in both of its two subnetworks. The output of G_1 are two residual images (skin lesion and segmentation mask) to be added to the expanded low resolution samples, provided by the previous pyramid level. This approach allows to enlarge images generated by the previous layer without lowering the resolution. Each following level has the same structure as the one employing G_1 . However, G_4 is used to provide residual images for both of the two last pyramid levels.

Fig. A.4 illustrates how G_3 adds noise in each sample, generating hair poorly, whereas G_4 does a great job improving the resolution and the realism of every image. As the image dimensions grow, target residual images of adjacent pyramid levels become more similar one another, allowing us to exploit the same GAN in more than one layer of the pyramid.

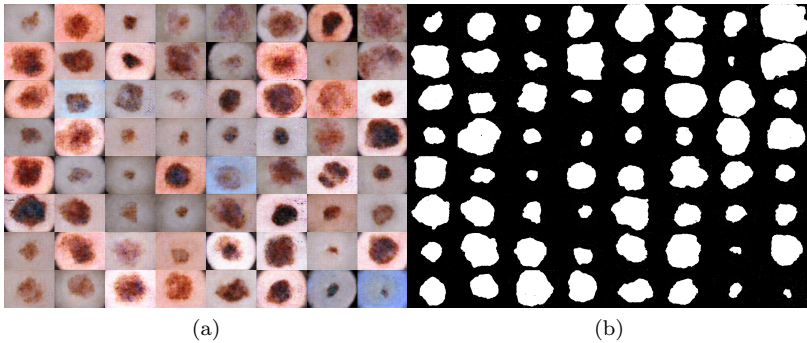


Figure A.1: DCGAN-generated skin lesion samples (a) and their segmentation masks (b), generated by a GAN trained with 500 dermoscopic samples.

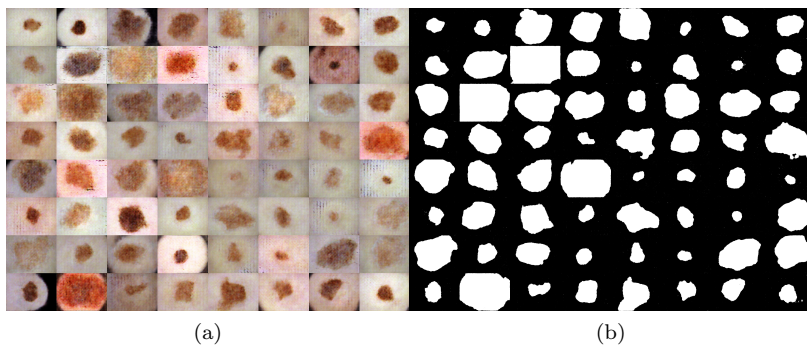


Figure A.2: DCGAN-generated skin lesion samples (a) and their segmentation masks (b), generated by a GAN trained with 1000 dermoscopic samples.

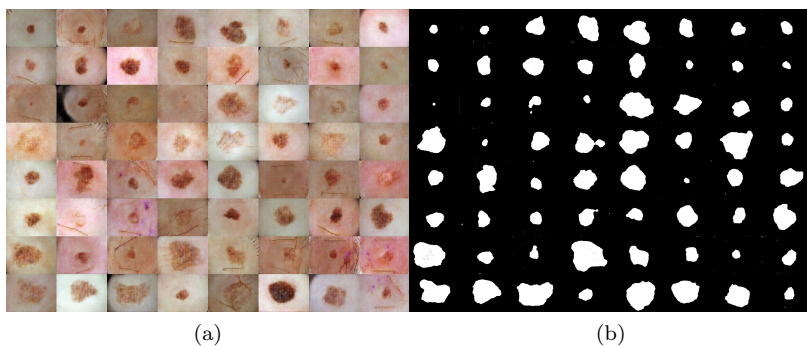


Figure A.3: DCGAN-generated skin lesion samples (a) and their segmentation masks (b), generated by a GAN trained with 1500 dermoscopic samples.

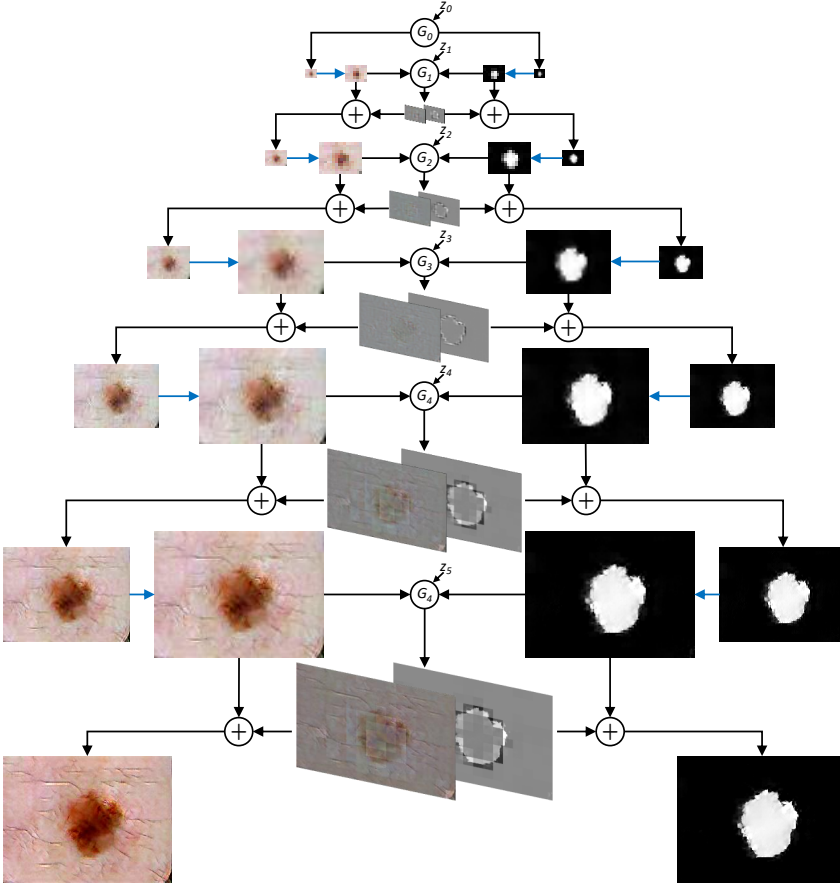


Figure A.4: A visual representation of our LAPGAN sampling process. For each level but the first one images are upsampled (blue arrows) and fed, together with a new source of noise, to a Convolutional Generative Adversarial Network, which serves to generate residual images. We add an extra step employing G_4 for a second time at the end of the process, in order to obtain 192×256 pixels samples.

Where Do Neural Networks Look?, 7

Additional examples of the data obtained after debasing ABCD properties are reported in Fig. A.5. When applying border removal, the thickness of the black line depends on the size of the lesion. This can be better appreciated by looking at the benign lesions (second half of the rows). When altering lesion diameter, the *Crop&Pad* transformation has higher effects on non-melanoma images.

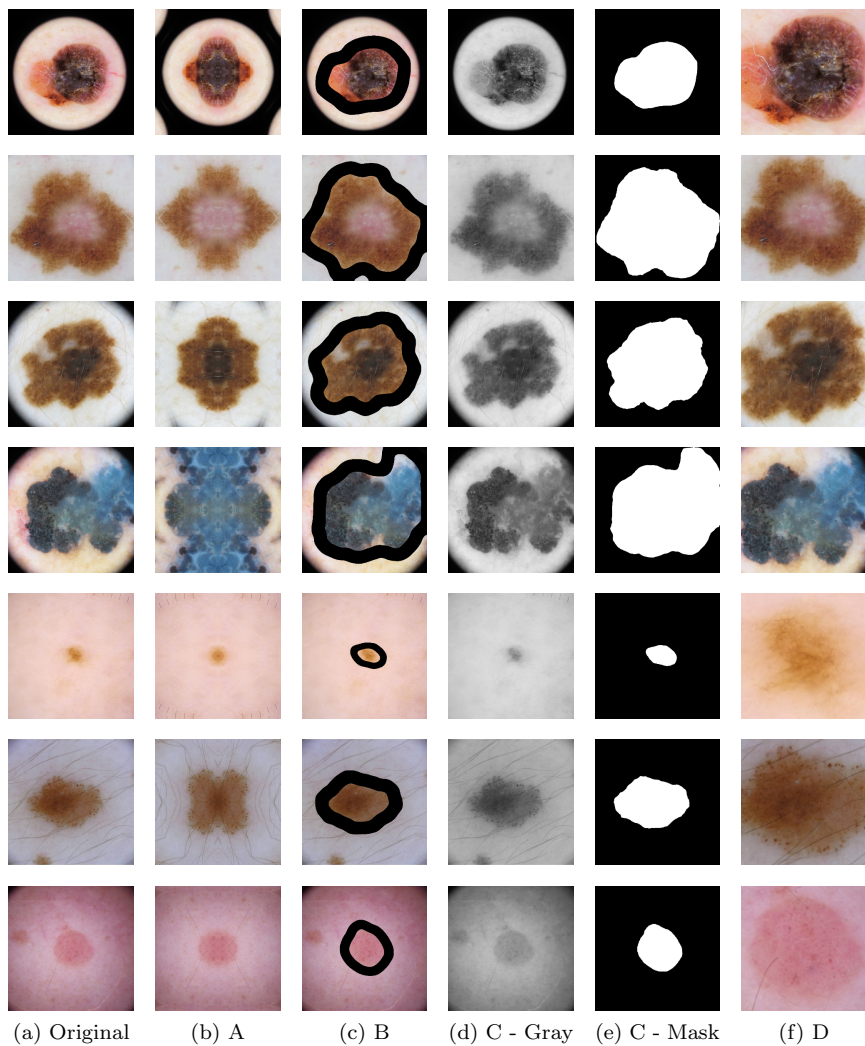


Figure A.5: Skin lesion image samples obtained from ISIC dataset after debasing different ABCD properties. The first half rows depict melanomas, while the others are generated from benign lesions.

Appendix B

List of Publications

This Section contains the list of research papers published during the Ph.D. period, alongside those accepted but yet to be released. Content and experimental results published in some of these papers have been included in this thesis, with explicit permission given by the other authors. Since the focus of this work is deep learning in dermoscopic image analysis, unrelated published papers are not discussed in this thesis.

- [1] Federico Pollastri, Federico Bolelli, Roberto Paredes, and Costantino Grana. Improving Skin Lesion Segmentation with Generative Adversarial Networks. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 442–443. IEEE, 2018. 10, 21, 36, 43
- [2] Federico Bolelli, Lorenzo Baraldi, Federico Pollastri, and Costantino Grana. A Hierarchical Quasi-Recurrent approach to Video Captioning. In *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, pages 162–167. IEEE, 2018.
- [3] Federico Pollastri, Federico Bolelli, Roberto Paredes, and Costantino Grana. Augmenting Data with GANs to Segment Melanoma Skin Lesions. In *Multimedia Tools and Applications Journal*. MTAP, Springer, 2019. 43
- [4] Stefano Allegretti, Federico Bolelli, Michele Cancilla, Federico Pollastri, Laura Canalini, and Costantino Grana. How

Does Connected Components Labeling with Decision Trees Perform on GPUs? In *International Conference on Computer Analysis of Images and Patterns*, pages 39–51. Springer, 2019.

- [5] Laura Canalini, [Federico Pollastri](#), Federico Bolelli, Michele Cancilla, Stefano Allegretti, and Costantino Grana. Skin Lesion Segmentation Ensemble with Diverse Training Strategies. In *International Conference on Computer Analysis of Images and Patterns*, pages 89–101. Springer, 2019. 53, 62, 78
- [6] Giulia Ligabue, [Federico Pollastri](#), Francesco Fontana, Marco Leonelli, Luciana Furci, Silvia Giovanella, Gaetano Alfano, Gianni Cappelli, Francesca Testa, Federico Bolelli, Costantino Grana, and Riccardo Magistroni. Evaluation of the Classification Accuracy of the Kidney Biopsy Direct Immunofluorescence through Convolutional Neural Networks. *Clinical Journal of the American Society of Nephrology*, 15(10):1445–1454, October 2020.
- [7] [Federico Pollastri](#), Juan Maroñas, Federico Bolelli, Giulia Ligabue, Roberto Paredes, Riccardo Magistroni, and Costantino Grana. Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1–8. IEEE, 2021. 51, 73
- [8] [Federico Pollastri](#), Mario Parreño, Juan Maroñas, Federico Bolelli, Roberto Paredes, Daniel Ramos, and Costantino Grana. A Deep Analysis on High Resolution Dermoscopic Image Classification. *IET Computer Vision*, pages 1–10, 2021. 82
- [9] Stefano Allegretti, Federico Bolelli, [Federico Pollastri](#), Sabrina Longhitano, Giovanni Pellacani, and Costantino Grana. Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1–8. IEEE, Jan 2021.
- [10] Cristian Mercadante, Marco Cipriano, Federico Bolelli, [Federico Pollastri](#), Alexandre Anesi, Costantino Grana, et al. A Cone Beam Computed Tomography Annotation Tool for Automatic Detection of the Inferior Alveolar Nerve Canal. In *16th International Conference on Computer Vision Theory and Applications-VISAPP 2021*, volume 4, pages 724–731. SciTePress, 2021.

- [11] Federico Pollastri, Marco Cipriano, Federico Bolelli, and Costantino Grana. Long-Range 3D Self-Attention for MRI Prostate Segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–1. IEEE, 2022.
- [12] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, and Costantino Grana. Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–1, 2022.
- [13] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Mattia Di Bartolomeo, Federico Pollastri, Arrigo Pellacani, Paolo Minafra, Alexandre Anesi, and Costantino Grana. Deep Segmentation of the MandibularCanal: a New 3D Annotated Dataset of CBCT Volumes. *IEEE Access*, 2022.
- [14] Dana Oniga, Barbara Cantalupo, Enzo Tartaglione, Daniele Perlo, Marco Grangetto, Marco Aldinucci, Federico Bolelli, Federico Pollastri, Michele Cancilla, Laura Canalini, Costantino Grana, Cristina Alcalde, Franco Alberto Cardillo, and Monica Florea. Applications of AI and HPC in the Health Domain. In Oliver Terzo and Jan Martinovic, editors, *HPC, Big Data, and AI Convergence Towards Exascale: Challenge and Vision*, chapter 11, pages 217–241. CRC Press, 2022.

Appendix C

Activities Carried Out During the PhD

Research activities carried out during the 3 years of PhD are reported in this chapter.

Participation to European Projects

- *H2020 - DEEPHEALTH*: a project funded under the call ICT-11-2018-2019 and currently on going. The DeepHealth project will provide HPC computing power at the service of biomedical applications; and apply Deep Learning (DL) techniques on large and complex biomedical datasets to support new and more efficient ways of diagnosis, monitoring and treatment of diseases. Among the other things, we are responsible for the development of the European Computer Vision Library (ECVL), one of the core elements of the project.

Teaching Activities

- Tutor for the “Fundamentals of Computer Science I” undergraduate course, at Università degli Studi di Modena e Reggio Emilia (2018-2021).

- Lecturer for the “Data Lab” courses, funded by Emilia Romagna and the European Social Fund and organized by CIS —Reggio Emilia—, IFOA —Reggio Emilia—, and Nuova Didactica —Modena— (2019).

Grants and Awards

- Third place (out of 64 research groups) at the international competition on skin lesion classification (2019 ISIC Challenge).
- Best paper award at the “18th International Conference on Computer Analysis of Images and Patterns” - CAIP.

Other Academic Services

Journals Reviewing

- ACM Transactions on Multimedia Computing, Communications, and Applications.
- IET Electronic Letters.
- IET Image Processing.
- IET Computer Vision.

Conferences Reviewing

- International Conference on Pattern Recognition.
- International Conference on Image Analysis and Processing.

Collaborations with Companies

- Collaboration with MediciMedical S.R.L. for the integration of state-of-the-art skin lesion classification algorithms into VIDIX acquisition software.

MSc thesis co-advisor

- Luca Benzi, “Deep learning techniques applied on medical images”, 2020.
- Laura Canalini, “Segmentation techniques analysis and testing using convolutional neural networks on skin lesion images”, 2019.

Attended Conferences, Schools, Courses, and Seminars

Conferences, Schools and Courses

- Imaging AI based Management of COVID-19 Webinar Series — On-line, 2020.
- International Conference on Pattern Recognition — ICPR, Milano, Italy, (Online) 2020.
- International Conference on Medical Image Computing and Computer Assisted Intervention — MICCAI, Shenzhen, China, 2019.
- International Conference on Computer Analysis of Images and Patterns - CAIP, Salerno, Italy, 2019.
- Academic English Workshop I — Silvia Cavalieri February, 2019.
- Academic English Workshop II — Silvia Cavalieri March, 2019.
- International Workshop on Computer Vision — IWCV, Modena, Italy, 2018.
- Medical Imaging Summer School — MISS, Favignana, Italy, 2018.

Seminars

- “About Time” — Prof. Arnold Smeulders September, 2020.
- “Explainable Machine Learning Models for Healthcare AI” — Carly Eckert September, 2020.

- “Empowering Citizens Leveraging their Data Power” — Prof. Marco Mellia May, 2020.
- “NVAITC Explained” — Frederic Pariente May, 2020
- “Computational Aspects of Deep Reinforcement Learning” — Iuri Frosio July, 2019.
- “Big Data in Neuroimaging” — Prof. Mark Jenkinson June, 2019.
- “Deep Learning-based Automatic Speech Recognition - A multimodal approach” — Prof. Leonardo Badino March, 2019.
- “Computer Graphics for Cultural Heritage Applications” — Roberto Scopigno March, 2019.
- “Deep Learning for Fault Prediction” — Prof. Roberto Paredes Palacios February, 2018.

Bibliography

- [1] 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,” *Dermoscopy Image Analysis*, pp. 97–129, 2015. 1
- [2] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, “Diagnostic accuracy of dermoscopy,” *The lancet oncology*, vol. 3, no. 3, pp. 159–165, 2002. 1
- [3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, Nov 2012. 1
- [4] F. Bolelli, L. Baraldi, and C. Grana, “A Hierarchical Quasi-Recurrent approach to Video Captioning,” in *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, pp. 162–167, Dec 2018. 1
- [5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1, 5, 28, 48
- [6] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *International Conference on Machine Learning*, pp. 6105–6114, 2019. 1, 6, 39, 48

- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009. 2, 6
- [8] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. E. Hopcroft, “Convergent Learning: Do different neural networks learn the same representations?,” in *FE@ NIPS*, pp. 196–212, 2015. 2
- [9] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” *arXiv preprint arXiv:1811.12231*, 2018. 2, 73
- [10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017. 3
- [11] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. 3
- [12] Y. Liu, L. Nie, L. Han, L. Zhang, and D. S. Rosenblum, “Action2Activity: Recognizing Complex Activities from Sensor Data,” in *Twenty-fourth international joint conference on artificial intelligence*, 2015. 3
- [13] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, “From action to activity: Sensor-based activity recognition,” *Neurocomputing*, vol. 181, pp. 108–115, 2016. 3
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012. 3, 69
- [15] Y. Yuan, M. Chao, and Y. Lo, “Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1876–1886, 2017. 3, 15, 16, 36

- [16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *Advances in neural information processing systems*, pp. 2672–2680, 2014. 4
- [17] T. Neff, C. Payer, D. Štern, and M. Urschler, “Generative Adversarial Network based Synthesis for Supervised Medical Image Segmentation,” in *OAGM & ARW Joint Workshop 2017 on “Vision, Automation & Robotics”*, Verlag der Technischen Universität Graz, 2017. 4
- [18] C. Baur, S. Albarqouni, and N. Navab, “MelanoGANs: High Resolution Skin Lesion Synthesis with GANs,” *arXiv preprint arXiv:1804.04338*, 2018. 4
- [19] G. Pellacani, C. Grana, and S. Seidenari, “Algorithmic reproduction of asymmetry and border cut-off parameters according to the ABCD rule for dermoscopy,” *Journal of the European Academy of Dermatology and Venereology*, vol. 20, no. 10, pp. 1214–1219, 2006. 4, 78
- [20] M. Zortea, E. Flores, and J. Scharcanski, “A simple weighted thresholding method for the segmentation of pigmented skin lesions in macroscopic images,” *Pattern Recognition*, vol. 64, pp. 92–104, 2017. 4
- [21] D. H. Chung and G. Sapiro, “Segmenting Skin Lesions with Partial-Differential-Equations-Based Image Processing Algorithms,” *IEEE transactions on Medical Imaging*, vol. 19, no. 7, pp. 763–767, 2000. 4
- [22] P. Rubegni, A. Ferrari, G. Cevenini, *et al.*, “Differentiation between pigmented Spitz naevus and melanoma by digital dermoscopy and stepwise logistic discriminant analysis,” *Melanoma research*, vol. 11, no. 1, pp. 37–44, 2001. 4
- [23] J. Gao, J. Zhang, and M. G. Fleming, “A novel multiresolution color image segmentation technique and its application to dermatoscopic image segmentation,” in *Proceedings 2000 International Conference on Image Processing*, vol. 3, pp. 408–411, IEEE, 2000. 4

- [24] P. Schmid, “Lesion Detection in Dermatoscopic Images Using Anisotropic Diffusion and Morphological Flooding,” in *Proceedings 1999 International Conference on Image Processing*, vol. 3, pp. 449–453, IEEE, 1999. 4
- [25] H. Ganster, P. Pinz, R. Rohrer, *et al.*, “Automated Melanoma Recognition,” *IEEE transactions on medical imaging*, vol. 20, no. 3, pp. 233–239, 2001. 4
- [26] M. H. Jafari, E. Nasr-Esfahani, N. Karimi, S. M. R. Soroushmehr, S. Samavi, and K. Najarian, “Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 6, pp. 1021–1030, 2017. 4
- [27] 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*, vol. 36, no. 4, pp. 994–1004, 2017. 4
- [28] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional Networks for Biomedical Image Segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015. 4, 19, 36
- [29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” *arXiv preprint arXiv:1412.7062*, 2014. 4, 34
- [30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018. 5, 28, 29, 31, 36, 78
- [31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017. 5, 32
- [32] R. Girshick, “Fast R-CNN,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015. 5

- [33] D. Cheng, R. Liao, S. Fidler, and R. Urtasun, “Darnet: Deep Active Ray Network for Building Segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7431–7439, 2019. 5
- [34] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, “Rethinking Pre-training and Self-training,” *Advances in neural information processing systems*, vol. 33, pp. 3833–3845, 2020. 5
- [35] X. Zhang, H. Xu, H. Mo, J. Tan, C. Yang, L. Wang, and W. Ren, “DCNAS: Densely Connected Neural Architecture Search for Semantic Image Segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13956–13967, 2021. 5
- [36] A. Tao, K. Sapra, and B. Catanzaro, “Hierarchical multi-scale attention for semantic segmentation,” *arXiv preprint arXiv:2005.10821*, 2020. 5
- [37] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” *IEEE transactions on pattern analysis and machine intelligence*, 2021. 5
- [38] Y. Bengio, P. Simard, P. Frasconi, *et al.*, “Learning Long-Term Dependencies with Gradient Descent is Difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994. 5
- [39] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *arXiv preprint arXiv:1502.03167*, 2015. 5, 11
- [40] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of Machine Learning Research*, pp. 249–256, 2010. 5
- [41] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, “Efficient BackProp,” in *Neural Networks: Tricks of the Trade*, pp. 9–48, Springer, 2012. 5

- [42] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500, 2017. 5
- [43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017. 5, 48
- [44] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, “Deep learning for image-based cancer detection and diagnosis - A survey,” *Pattern Recognition*, vol. 83, pp. 134–149, 2018. 6
- [45] K. He, R. Girshick, and P. Dollár, “Rethinking ImageNet Pre-Training,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4918–4927, 2019. 6
- [46] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding Transfer Learning for Medical Imaging,” in *Advances in Neural Information Processing Systems*, pp. 3342–3352, 2019. 6
- [47] D. S. Rigel, J. Russak, and R. Friedman, “The Evolution of Melanoma Diagnosis: 25 Years Beyond the ABCDs,” *CA: A Cancer Journal for Clinicians*, vol. 60, no. 5, pp. 301–316, 2010. 6, 73, 78
- [48] B. Harangi, “Skin lesion classification with ensembles of deep convolutional neural networks,” *Journal of Biomedical Informatics*, vol. 86, pp. 25–32, 2018. 6
- [49] M. A. Al-Masni, D.-H. Kim, and T.-S. Kim, “Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification,” *Computer Methods and Programs in Biomedicine*, vol. 190, p. 105351, 2020. 6
- [50] E. Valle, M. Fornaciali, A. Menegola, J. Tavares, F. Vasques Bittencourt, L. T. Li, and S. Avila, “Data, depth, and design: Learning reliable models for skin lesion analysis,” *Neurocomputing*, vol. 383, pp. 303–313, 2020. 6

- [51] L. Ballerini, X. Li, R. B. Fisher, and J. Rees, "A Query-by-Example Content-Based Image Retrieval System of Non-melanoma Skin Lesions," in *MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support*, pp. 31–38, Springer, 2009. 6
- [52] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967. 6
- [53] A. Baldi, R. Murace, E. Dragonetti, M. Manganaro, O. Guerra, S. Bizzi, and L. Galli, "Definition of an automated Content-Based Image Retrieval (CBIR) system for the comparison of dermoscopic images of pigmented skin lesions," *BioMedical Engineering OnLine*, vol. 8, no. 1, p. 18, 2009. 6
- [54] W. Jiji and J. Durai Raj, "Content-based image retrieval in dermatology using intelligent technique," *IET Image Processing*, vol. 9, pp. 306–317, 2014. 6
- [55] M. Rahman, N. Alpaslan, and P. Bhattacharya, "Developing a retrieval based diagnostic aid for automated melanoma recognition of dermoscopic images," in *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–7, 2016. 7
- [56] K. Belattar, S. Mostefai, and A. Draa, "Intelligent Content-Based Dermoscopic Image Retrieval with Relevance Feedback for Computer-Aided Melanoma Diagnosis," *Journal of Information Technology Research*, vol. 10, pp. 85–108, 2017. 7
- [57] X. Pu, Y. Li, H. Qiu, and Y. Sun, "Deep Semantics-Preserving Hashing Based Skin Lesion Image Retrieval," in *Advances in Neural Networks - ISNN 2017*, pp. 282–289, Springer, 2017. 7, 64, 68, 69, 70
- [58] H. Yang, K. Lin, and C. Chen, "Supervised Learning of Semantics-Preserving Hash via Deep Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 437–451, 2018. 7
- [59] G. Maguolo and L. Nanni, "A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images," *arXiv preprint 2004.12823*, 2020. 7, 77

- [60] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, *et al.*, “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans,” *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021. 7, 90
- [61] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, “(De)Constructing Bias on Skin Lesion Datasets,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–9, 2019. 8, 73, 87, 88
- [62] 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*, vol. 111, pp. 30–37, 2019. 8
- [63] B. Zhou, D. Bau, A. Oliva, and A. Torralba, “Interpreting Deep Visual Representations via Network Dissection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2131–2145, 2019. 8
- [64] C. Barata, M. E. Celebi, and J. S. Marques, “Explainable Skin Lesion Diagnosis Using Taxonomies,” *Pattern Recognition*, vol. 110, p. 107413, 2020. 8
- [65] R. R. Selvaraju, M. Cogswell, A. Das, *et al.*, “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. 8, 28, 55, 62, 78, 93
- [66] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, “Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020. 8
- [67] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, “Improving Skin Lesion Segmentation with Generative Adversarial Networks,” in *2018*

IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pp. 442–443, IEEE, 2018. 10, 21, 36, 43

- [68] D. Mishkin and J. Matas, “All you need is a good init,” in *International Conference on Learning Representations (ICLR) 2016*, 2016. 10
- [69] J. T. Springenberg, “Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks,” *arXiv preprint arXiv:1511.06390*, 2015. 11
- [70] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled Samples Generated by GAN Improve the Person Re-Identification Baseline in Vitro,” *arXiv preprint arXiv:1701.07717*, vol. 3, 2017. 11, 30
- [71] A. Antoniou, A. Storkey, and H. Edwards, “Data Augmentation Generative Adversarial Networks,” *arXiv preprint arXiv:1711.04340*, 2017. 11
- [72] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *arXiv preprint arXiv:1511.06434*, 2015. 11
- [73] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier Nonlinearities Improve Neural Network Acoustic Models,” in *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing (WD-LASL 2013)*, 2013. 11
- [74] E. L. Denton, S. Chintala, R. Fergus, *et al.*, “Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks,” in *Advances in neural information processing systems*, pp. 1486–1494, 2015. 12
- [75] A. H. Lipkus, “A proof of the triangle inequality for the Tanimoto distance,” *Journal of Mathematical Chemistry*, vol. 26, pp. 263–265, Oct 1999. 16
- [76] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv preprint arXiv:1412.6980*, 2014. 18, 28, 29, 82

- [77] N. Codella, D. Gutman, Celebi, *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018. 18, 22, 41, 63, 66, 75
- [78] F. Bolelli, L. Baraldi, M. Cancilla, and C. Grana, “Connected Components Labeling on DRAGs,” in *International Conference on Pattern Recognition*, 2018. 20, 33
- [79] F. Bolelli, M. Cancilla, and C. Grana, “Two More Strategies to Speed Up Connected Components Labeling Algorithms,” in *International Conference on Image Analysis and Processing*, pp. 48–58, Springer, 2017. 20, 33
- [80] 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*, vol. 5, 2018. 28, 66
- [81] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” in *International Conference on Learning Representations*, 2018. 30
- [82] W. Liu, A. Rabinovich, and A. C. Berg, “ParseNet: Looking Wider to See Better,” *arXiv preprint arXiv:1506.04579*, 2015. 31
- [83] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, “Microsoft COCO: Common Objects in Context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014. 32
- [84] A. Krogh and J. Vedelsby, “Neural Network Ensembles, Cross Validation, and Active Learning,” in *Advances in neural information processing systems*, pp. 231–238, 1995. 33
- [85] A. Chandra and X. Yao, “Evolving Hybrid Ensembles of Learning Machines for Better Generalisation,” *Neurocomputing*, vol. 69, no. 7-9, pp. 686–700, 2006. 33
- [86] Y. Xue, T. Xu, and X. Huang, “Adversarial learning with multi-scale loss for skin lesion segmentation,” in *2018 IEEE 15th International*

- Symposium on Biomedical Imaging (ISBI 2018)*, pp. 859–863, IEEE, 2018. 36
- [87] L. Bi, J. Kim, E. Ahn, A. Kumar, D. Feng, and M. Fulham, “Step-wise integration of deep class-specific learning for dermoscopic image segmentation,” *Pattern Recognition*, vol. 85, pp. 78–89, 2019. 36
 - [88] X. Li, L. Yu, C.-W. Fu, and P.-A. Heng, “Deeply Supervised Rotation Equivariant Network for Lesion Segmentation in Dermoscopy Images,” in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pp. 235–243, Springer, 2018. 36
 - [89] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. 36
 - [90] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio, “The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11–19, 2017. 36
 - [91] L. Canalini and F. Pollastri, *Source Code*. available at https://github.com/PollastriFederico/skin_lesion_segmentation_ensemble. 37
 - [92] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1. MIT press Cambridge, 2016. 39, 48
 - [93] C. Mercadante, M. Cipriano, F. Bolelli, F. Pollastri, M. Di Bartolomeo, A. Anesi, and C. Grana, “A Cone Beam Computed Tomography Annotation Tool for Automatic Detection of the Inferior Alveolar Nerve Canal,” in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 1–8, Feb 2021. 39
 - [94] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, A. C. Halpern, S. Puig, and J. Malveyh, “BCN20000:

Dermoscopic Lesions in the Wild,” *arXiv preprint arXiv:1908.02288*, pp. 1–3, 2019. 41, 66, 75

- [95] J. D. Whited and J. M. Grichnik, “Does This Patient Have a Mole or a Melanoma?,” *Jama*, vol. 279, no. 9, pp. 696–701, 1998. 42
- [96] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, “Data Augmentation for Skin Lesion Analysis,” in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pp. 303–311, Springer, 2018. 43
- [97] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, “Augmenting Data with GANs to Segment Melanoma Skin Lesions,” in *Multimedia Tools and Applications Journal*, MTAP, Springer, 2019. 43
- [98] T. DeVries and G. W. Taylor, “Improved Regularization of Convolutional Neural Networks with Cutout,” *arXiv preprint arXiv:1708.04552*, pp. 1–8, 2017. 44
- [99] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. 48
- [100] D. J. C. MacKay, “The Evidence Framework Applied to Classification Networks,” *Neural Computation*, vol. 4, no. 5, pp. 720–736, 1992. 49
- [101] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 6402–6413, Curran Associates, Inc., 2017. 49, 53
- [102] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. 49, 52
- [103] H.-C. Kim and Z. Ghahramani, “Bayesian Classifier Combination,” in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (N. D. Lawrence and M. Girolami, eds.),

- vol. 22 of *Proceedings of Machine Learning Research*, (La Palma, Canary Islands), pp. 619–627, PMLR, Apr 2012. 49
- [104] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, “Skin Lesion Classification Using CNNs With Patch-Based Attention and Diagnosis-Guided Loss Weighting,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 495–503, 2019. 51
 - [105] F. Pollastri, J. Maroñas, F. Bolelli, G. Ligabue, R. Paredes, R. Magistroni, and C. Grana, “Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 1–8, IEEE, 2021. 51, 73
 - [106] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. 51, 53
 - [107] M. H. deGroot and S. E. Fienberg, “The Comparison and Evaluation of Forecasters,” *The Statistician*, vol. 32, pp. 12–22, 1983. 52
 - [108] L. Canalini, F. Pollastri, F. Bolelli, M. Cancilla, S. Allegretti, and C. Grana, “Skin Lesion Segmentation Ensemble with Diverse Training Strategies,” in *International Conference on Computer Analysis of Images and Patterns*, pp. 89–101, Springer, 2019. 53, 62, 78
 - [109] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, “Averaging Weights Leads to Wider Optima and Better Generalization,” *arXiv preprint arXiv:1803.05407*, pp. 1–12, 2018. 53
 - [110] A. Kumar *et al.*, “Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings,” in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 2805–2814, Jul 2018. 53
 - [111] J. Maroñas, R. Paredes, and D. Ramos, “Calibration of Deep Probabilistic Models with Decoupled Bayesian Neural Networks,” *Neurocomputing*, vol. 407, pp. 194–205, 2020. 53

- [112] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data,” *MethodsX*, p. 100864, 2020. 55, 57, 59
- [113] M. Cancilla, L. Canalini, F. Bolelli, S. Allegretti, S. Carrión, R. Paredes, J. A. Gómez, S. Leo, M. E. Piras, L. Pireddu, A. Badouh, S. Marco-Sola, L. Alvarez, M. Moreto, and C. Grana, “The Deep-Health Toolkit: A Unified Framework to Boost Biomedical Applications,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 1–8, IEEE, Jan 2021. 57, 78
- [114] *ISIC 2019 Leaderboard*. available at <https://challenge.isic-archive.com/leaderboards/2019>. 59
- [115] F. Pollastri, M. Parreño, J. Maroñas, and F. Bolelli, *Source Code*. available at <https://github.com/PollastriFederico/ISIC2019>. 60
- [116] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, *et al.*, “Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study,” *The Lancet Oncology*, vol. 20, no. 7, pp. 938–947, 2019. 61
- [117] S. Dreiseitl and M. Binder, “Do physicians value decision support? A look at the effect of decision support systems on physician opinion,” *Artificial Intelligence in Medicine*, vol. 33, no. 1, pp. 25–30, 2005. 61
- [118] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019. 62
- [119] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, “Large Scale Online Learning of Image Similarity Through Ranking,” *Journal of Machine Learning Research - JMLR*, vol. 11, pp. 11–14, 2009. 64
- [120] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015. 65, 68

- [121] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021. 73
- [122] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut Learning in Deep Neural Networks,” *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020. 73
- [123] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Choussakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, *et al.*, “A patient-centric dataset of images and metadata for identifying melanomas using clinical context,” *Scientific data*, vol. 8, no. 1, pp. 1–8, 2021. 76
- [124] N. F. Lattoofi, I. F. Al-Sharuee, M. Y. Kamil, A. H. Obaid, A. A. Mahidi, A. A. Omar, *et al.*, “Melanoma Skin Cancer Detection Based on ABCD Rule,” in *2019 First International Conference of Computer and Applied Sciences (CAS)*, pp. 154–157, IEEE, 2019. 78
- [125] N. C. Lynn and Z. M. Kyu, “Segmentation and Classification of Skin Cancer Melanoma from Skin Lesion Images,” in *2017 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, pp. 117–122, IEEE, 2017. 79
- [126] Q. Ha, B. Liu, and F. Liu, “Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge,” *arXiv preprint 2010.05351*, 2020. 81
- [127] J. Weber, *True duplicates in ISIC 2020 dataset*. available at <https://www.kaggle.com/c/siim-isic-melanoma-classification/discussion/161943>. 81
- [128] C. Gaudy-Marqueste, Y. Wazaefi, Y. Bruneu, R. Triller, L. Thomas, G. Pellacani, J. Malvehy, M.-F. Avril, S. Monestier, M.-A. Richard, *et al.*, “Ugly Duckling Sign as a Major Factor of Efficiency in Melanoma Detection,” *JAMA dermatology*, vol. 153, no. 4, pp. 279–284, 2017. 82
- [129] F. Pollastri, M. Parreño, J. Maroñas, F. Bolelli, R. Paredes, D. Ramos, and C. Grana, “A Deep Analysis on High Resolution

Dermoscopic Image Classification,” *IET Computer Vision*, pp. 1–10, 2021. 82

- [130] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” *Neurocomputing*, vol. 312, pp. 135–153, 2018. 87
- [131] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Universal Domain Adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, June 2019. 87
- [132] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, “Unsupervised domain adaptation for medical imaging segmentation with self-ensembling,” *NeuroImage*, vol. 194, pp. 1–11, 2019. 87
- [133] E. Claridge, P. Hall, M. Keefe, and J. Allen, “Shape analysis for classification of malignant melanoma,” *Journal of Biomedical Engineering*, vol. 14, no. 3, pp. 229–234, 1992. 88, 93
- [134] E. M. Senan and M. E. Jadhav, “Analysis of Dermoscopy Images by Using ABCD Rule for Early Detection of Skin Cancer,” *Global Transitions Proceedings*, 2021. 88, 93
- [135] A. Brú, S. Albertos, J. L. Subiza, J. L. García-Asenjo, and I. Brú, “The Universal Dynamics of Tumor Growth,” *Biophysical journal*, vol. 85, no. 5, pp. 2948–2961, 2003. 93
- [136] G. Capdehourat, A. Corez, A. Bazzano, R. Alonso, and P. Musé, “Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions,” *Pattern Recognition Letters*, vol. 32, no. 16, pp. 2187–2196, 2011. Advances in Theory and Applications of Pattern Recognition, Image Processing and Computer Vision. 93