DEEP LEARNING FOR SKIN CANCER DIAGNOSIS WITH HIERARCHICAL ARCHITECTURES

Catarina Barata and Jorge S. Marques

Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal

ABSTRACT

Skin lesions are organized in a hierarchical way, which is taken into account by dermatologists when diagnosing them. However, automatic systems do not make use of this information, performing the diagnosis in a one-vs-all approach, where all types of lesions are considered. In this paper we propose to mimic the medical strategy and train a deep-learning architecture to perform a hierarchical diagnosis. Our results highlight the benefits of addressing the classification of dermoscopy images in a structured way. Additionally, we provide an extensive evaluation of criteria that must be taken into account in the development of diagnostic systems based on deep learning.

***Index Terms*—** Skin Cancer, Hierarchical Classification, Deep Learning, Dermoscopy

1. INTRODUCTION

Skin cancer is one of the most common types of cancer worldwide, accounting for approximately one third of all the diagnoses. The overwhelming increase in its incidence rates, particularly of melanoma that has grown over 300% from 1990 to 2018 just in the US [1], has raised the attention of researchers. In particular, there is a focus on the development of methods for the automatic diagnosis of dermoscopy images [2].

Although dermoscopy image analysis has been an active topic of research for more than twenty years, the last couple of years have seen a significant increase in the number of published works [2]. Such interest has been mainly encouraged by the release of public dermoscopy datasets, such as PH^2 [3] and the ISIC challenges [4, 5]. Moreover, the deep learning revolution [6] has also played a role, with the proposal of increasingly deeper and better convolutional architectures (CNN) and the release of open source software tools. Deep learning and small datasets, such as the dermoscopy ones, are antagonists, meaning that it is not reasonable to train CNN architectures from scratch to tackle the problem of skin cancer. However, the availability of pre-trained networks, which may be used for transfer learning either as *feature extractors* or as a starting point for *fine-tuning* to the skin cancer problem, has fostered the release of several works based on this methodology [7].

The most recent public datasets have extended the traditional melanoma/benign problem using only melanocytic lesions, to a multi-class one where non-melanocytic lesions have been added (*e.g.*, ISIC 2017 [5]). Several methods have treated this problem has a one-vs-all approach, where the network tries to distinguish between all of the classes in the same decision layer. But, dermatologists divide this task into a hierarchical method: first they distinguish between melanocytic/non-melanocytic and only then they perform the final diagnosis [8].

Thus, it is possible to wonder if there is any benefit in mimicking the medical diagnosis, and train hierarchical networks. This paper shows that it is better to use hierarchical networks. Additionally, we conduct several experiments that shed some light on the following points: i) importance of color normalization and lesion segmentation; ii) performance of transfer learning strategies; and iii) comparison of evaluation metrics. To the best of our knowledge this is the first work that explores the hierarchical organization of skin lesions and simultaneously investigates points i), ii), and iii).

The remaining of the paper is organized as follows. Section 2 gives an overview of CNN architectures in skin cancer diagnosis, Section 3 introduces the hierarchical architectures, and Section 4 describes the experimental setup. Section 5 presents the results and Section 6 concludes the paper.

2. CNNS IN DERMOSCOPY IMAGE ANALYSIS

For the past years, CNNs have been used in dermoscopy image analysis. One of the first works is that of Codella et al. [9] where the Caffe architecture was used as a feature extractor. Esteva et al. [10] trained an Inception network from scratch using a very large private dataset of both clinical and dermoscopy images, showing that it was possible to achieve a performance similar to a human expert. However, training a CNN from scratch to diagnose skin cancer is usually infeasible due to the reduced size of the datasets (*e.g.*, the dataset from the 2017 challenge contained only 2000 images). Therefore, most works have either used pre-trained CNNs as feature extractors or have fine-tuned them for this problem [7].

This work was supported by the FCT project and plurianual funding: [PTDC/EEIPRO/0426/2014], [UID/EEA/50009/2019]. The Titan Xp used for this research was donated by the NVIDIA Corporation.

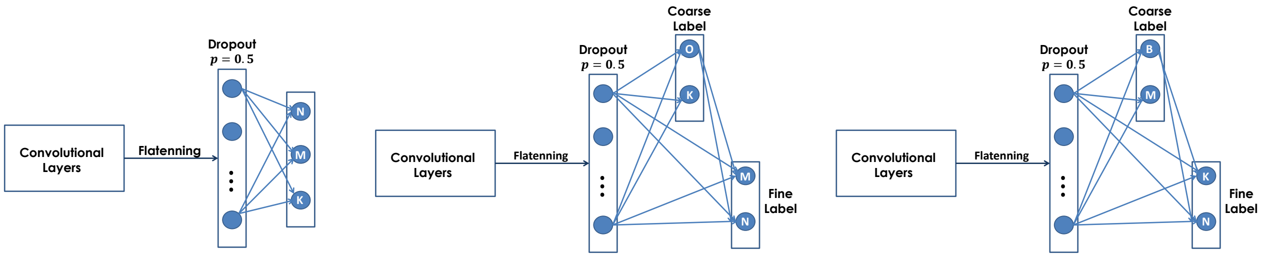

Fig. 1: Classification strategies: multi-class (left), hierarchical melanocytic-non melanocytic ($hier_1$-mid), and hierarchical malignant-benign ($hier_2$-right). Here, $\mathbf{o}$ identifies the melanocytic class and $\mathbf{p}$ stands for the dropout probability.

The use of CNNs was extensively observed in the 2017 [5] and 2018 [1] ISIC challenges. While in 2017 most participants showed a preference for ResNet, Inception, and ResNext architectures, in 2018 the use of deeper and more complex architectures, such as DenseNet and PNASNet, was also observed . Another difference in the two challenges is the use of ensembles of CNNs in 2018, which had already been pointed out by the challenge organizers as a way to improve the results [5]. Recently, several ensemble techniques have been proposed [11, 12], with promising results.

Some authors have devoted their work to studying specific aspects of the CNN that may improve the classification results of dermoscopy images. In particular, great importance has been given to the identification of suitable data augmentation strategies that may help dealing with the limited amount of available data [13, 14]. Additionally, attention has been paid to the comparison between transfer learning with and without fine-tuning [15], performing data augmentation on the test set [14], and other relevant criteria (*e.g*, image size and selected architecture) [11].

Although a hierarchical classification was investigated before using hand-crafted features [16], to the best of our knowledge, the application of this idea to CNNs has been poorly investigated in the dermoscopy field. The exception is the work of Demyanov et al. [17], which uses both clinical and dermoscopy images to train a ResNet-50 using a tree-loss function. This dataset is significantly different from the one used in our work, which contains only dermoscopy images. Moreover, we propose a simpler approach to impose hieararchy in our classification procedure.

3. HIERARCHICAL CNN

Dermoscopy lesions are categorized in a hierarchical way, where the lesions are firstly grouped in melanocytic or non-melanocytic, according to their origin, and only then diagnosed into a more fine category [8]. Although this hierarchy is well know in the literature, an evaluation of CNN architectures that perform a structured classification is still missing in the literature.

We address this problem and compare three classification strategies: one based on a multi-class formulation (see Fig. 1 (left)) and two based on hierarchical classification (see Fig. 1 (mid and right)). Our dataset contains examples of non-melanocytic lesions (seborrheic keratosis-**K**) and melanocytic lesions (melanoma-**M** and Nevi-**N**).

With respect to the hierarchical strategies, we aim to infer it is better to: i) mimic dermatologists and first discriminate between non-melanocytic (**K**) and melanocytic lesions (**M** and **N**) - $hier_1$; or ii) to first discriminate between malignant (**M**) and benign lesions (**K** and **N**) - $hier_2$.

4. EXPERIMENTAL SETUP

This section describes the experimental evaluation of the strategies proposed in Section 3. Additionally, we also assess the role of several factors that may influence the performance of deep neural networks. In the following sections we identify key aspects that are studied in the paper.

4.1. Dataset

For many years, the works devoted to skin cancer diagnosis used relatively small datasets, which usually comprised only examples of melanocytic lesions. Recently, the ISIC project started to release increasingly larger and more complex datasets associated with conference challenges. The challenges' datasets are particularly relevant, since they allow a fair comparison between methods and their performances. Therefore, in this work we will use the ISIC 2017-ISBI set [5], which is divided into training (2000 images), validation (150 images), and test (600 images) sets. The task of this challenge was to diagnose three classes of lesions: **M**, **K**, and **N**. Contrary to several of the challenge competitors, we will not augment the training set with external data, as we are interested in assessing how to make the most of a dataset, even if limited, to efficiently train deep learning architectures. Moreover, we want to ensure that our results are reproducible.

4.2. Pre-processing

It may be useful to perform several transformations to dermoscopy images before feeding them to a CNN. In this work

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

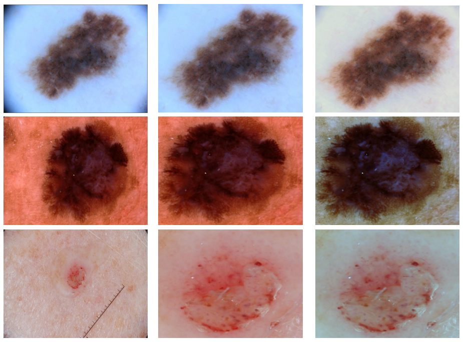

Fig. 2: Examples of pre-processed images: original (1st column); segmented and cropped (2nd column); normalized (3rd column).

we will focus in two types of transformations: lesion segmentation and color normalization.

Lesion segmentation corresponds to the separation between the lesion and the surrounding skin. In our experiments this will amount to cropping the original dermoscopy image with a tight bounding box around the lesion (see Fig. 2, 2nd column). Although the role of lesion segmentation is still an open issue in dermoscopy image analysis [7], it is important to understand how it influences the performance of CNN architectures.

Color normalization allows us to correct the colors of the dermoscopy images and reduce the variability introduced by the acquisition setup, as exemplified in Fig. 2, 3rd column. Similarly to the top classified of the ISIC-2017 challenge [18] and several of the participants of the ISIC-2018 challenge, we apply the color normalization strategy proposed in [19] to correct the image colors using its statistics. We set the value of $p = 6$.

After applying the aforementioned transformations, all of the images were resized to 299×299.

4.3. Network Training

Due to the reduced size of the training set we will use the DenseNet-161 architecture pre-trained on the ImageNet dataset [20], comparing two approaches: *feature extractor* vs *fine-tuning*. In the *feature extractor* learning approach we will freeze all the layers except the decision one(s), which will be trained for our problem, while in the *fine-tuning* case the pre-trained weights will be used as a soft initialization.

All of the models will be trained using the Adam Optimizer and a mini-batch approach, with a batch size of 5. The starting learning rate η will be $\eta = 0.005$ for transfer learning and $\eta = 10^{-5}$ for fine-tuning, with a decay rate of 0.5 for every 40 epochs. Cross-entropy is the selected loss function.

4.4. Generalization

It is crucial to train deep learning architectures that generalize well to new images. In this work we will rely on two strategies. The first one is based on online data augmentation, which consists of randomly flipping, rotating, cropping, and altering the colors of the training images in each epoch. We have picked this particular combination of transformations because they have been shown to improve the results of CNNs [13, 14]. Although online augmentation does not increase the size of the training set, it guarantees that the network ”sees” a different version of the same image between epochs, which reduces the probability of the network memorizing it and improves the generalization.

The other strategy is based on the use of dropout [21]. In particular, we will apply dropout with 50% probability, before the decision layer(s), as shown in Fig. 1.

4.5. Unbalanced Data

The training set used in this work is very unbalanced, with the following proportions: 18.7% **M**, 12.9% **K**, and 68.6% **N**. Popular approaches to deal with this issue are to artificially augment the less frequent classes, to assign different weights to the classes in the cost function, or to combine the previous two.

In this work we will resort to weighting the cross-entropy losses of the training examples. In particular, we will assign the class weights based on their distribution:

$$w_c = \frac{\#N}{\#N_c}, \tag{1}$$

where $\#N$ is the size of the training set and $\#N_c$ is the number of training elements form class $c \in \{\mathbf{M}, \mathbf{K}, \mathbf{N}\}$.

4.6. Evaluation Metrics

Finding the appropriate metrics to evaluate and compare the performance of classification systems is a challenging task. The metric used to rank the participants in the ISIC-2017 challenge was the average area under the curve (AUC) for the **M** and **K** diagnosis [5]. Thus, we will also apply this metric to evaluate the performance of the tested model configurations.

Although, AUC is a suitable metric to compare models, it is difficult to infer the performance of the model for each of the classes solely by inspecting its value. In the ISIC-2018 challenge, the ranking procedure was changed to be based on the balanced accuracy metric ($BACC$), which averages the recall (Re) values of all the class

$$Re = \frac{\#TP_c}{\#N_c}, \tag{2}$$

where TP_c is the number of true positives, *i.e*, the number of correctly classified examples from class c.

5. RESULTS

The experimental framework described in Section 4 was implement using Tensorflow and one Titan Xp GPU. Overall,

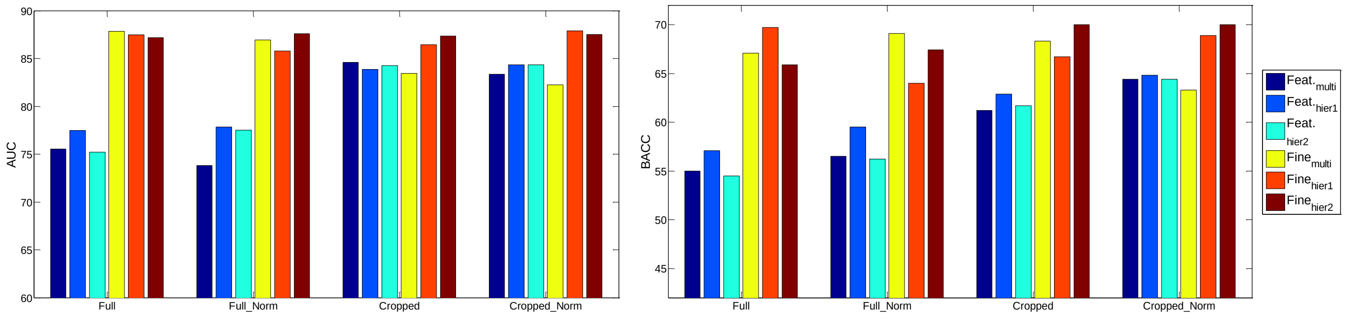

Fig. 3: Performance results for the test set: .

the experiments amounted to training and evaluating 24 different network architectures. The architectures were trained for 500 epochs using the 2000 training images, and validated every 10 epochs using the validation set. Figure 3 summarizes the AUC and $BACC$ scores for the test set.

These results yield relevant information. First, the use of a hierarchical classification strategy seems to lead to better overall results than using a traditional multi-class approach. Such results are observed both for the *feature extraction* (range of blues) and *fine-tuning* strategies (hot color bars), using any type of image pre-processing. As expected, fine-tuning DenseNet-161 to our problem leads to better experimental results both in terms of AUC and $BACC$. Interestingly, this improvement in more notorious in the AUC scores of the architectures trained with the full image (1st and 2nd sets of bars), while for the cropped images (3rd and 4th sets of bars) it seems that fine-tuning even degrades the performance of the multi-class architectures. However, when one inspects the $BACC$ scores, it is clear that fine-tuning leads to significant improvements in all of the cases, suggesting that the evaluation of a model must take into account more than one metric.

Cropped images seem to convey more discriminant information, specially when combined with the hierarchical architectures. In particular, the use of cropped images seems to be more suitable to diagnose melanomas, since the Re_M increases, as shown in Table 1. The scores shown in this table were obtained using the hierarchical architecture $hier_2$, *i.e.*, first discriminate between malignant and benign lesions and then between types of benign lesions. Contrary to what was expected, since $hier_1$ (orange bars) is the methodology used by dermatologists, $hier_2$ (red bars) seems to be the one that leads to the best results for most of the configurations. Such finding may be explained by the difficulty in diagnosing melanomas when compared with other types of skin lesions. This is a promising result that must be further investigated with a dataset that contains other types of malignant lesions, such as basal cell carcinomas [8].

Regarding the use of color normalization, it seems to lead to a marginal improvement in the AUC scores and to similar $BACC$ for the cropped images. However, when we take a closer look at the Re values for the different classes we observe that they are significantly different, evidencing again the importance of considering more than one metric to evaluate a classification system.

Table 1: Best performance scores.

Image Type	Re_M	Re_K	Re_N	AUC	$BACC$
Full	44.4%	70.0%	83.4%	87.2%	65.9%
Full Norm.	46.1%	71.1%	**85.0%**	**87.6%**	67.4%
Cropped	50.0%	**76.7%**	83.3%	87.4%	**70.0%**
Cropped Norm.	**59.8%**	71.1%	79.2%	87.5%	**70.0%**

We have compared our results with those of the ISIC challenge [5]. Our scores rank in the 70th percentile regarding the AUC metric, meaning that the hierarchical approach would rank above 7th position in the leaderboard. Regarding $BACC$, we have only compared our scores for the melanoma and keratosis classes, since these are the only Re available to the public). In this case, our hierarchical formulation would rank in the 90th percentile, with a $BACC = 65.5\%$. These are promising results, especially if one takes into account that we have used simple regularization techniques (dropout and online data augmentation) an no external data, to train our networks and prevent overfitting.

6. CONCLUSIONS

This paper explores the hierarchical organization of skin lesions, in order to develop a deep learning system that performs a structured classification. Additionally, we performed comparative studies on the importance of lesion segmentation, color normalization, and evaluation metrics.

Our results show that a structured classification based on a distinction between malignant and benign lesions, followed by the diagnosis of the latter in different classes leads to better results, when combined with segmented lesions. Color normalization also improves the results, but plays a minor role. Finally, we have also showed that our approach compares favorably with other state-of-the-art methods.

Future work should focus on validating these results on a larger dataset that comprises more classes of non-melanocytic lesions.

7. REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018,” *CA: a cancer journal for clinicians*, vol. 68, pp. 7–30, 2018.

[2] S. Pathan, K. G. Prabhu, and P. C. S., “Techniques and algorithms for computer aided diagnosis of pigmented skin lesions - a review,” *Biomedical Signal Processing and Control*, vol. 39, pp. 237–262, 2018.

[3] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, “PH2: A dermoscopic image database for research and benchmarking,” in *IEEE EMBC 2013*, 2013, pp. 5437–5440.

[4] D. Gutman, N. C. F. Codella, M. E. Celebi, and et al., “Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic),” *arXiv preprint arXiv:1605.01397*, 2016.

[5] N. C. F. Codella, D. Gutman, M. E. Celebi, and 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 *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE, 2018, pp. 168–172.

[6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] C. Barata, M. E. Celebi, and J. S. Marques, “A survey of feature extraction in dermoscopy image analysis of skin cancer,” *IEEE Journal of Biomedical and Health Informatics*, 2018.

[8] G. Argenziano, H P. Soyer, V. De Giorgi, and et al., *Interactive Atlas of Dermoscopy*, EDRA Medical Publishing & New Media, 2000.

[9] N. C. F Codella, J. Cai, M. Abedini, and et al., “Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images,” in *MLMI 2015*, 2015, pp. 118–126.

[10] A. Esteva, B. Kuprel, R. A Novoa, and et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115–118, 2017.

[11] E. Valle, M. Fornaciali, A. Menegola, and et al., “Data, depth, and design: learning reliable models for melanoma screening,” *arXiv preprint arXiv:1711.00441*, 2017.

[12] B. Harangi, “Skin lesion classification with ensembles of deep convolutional neural networks,” *Journal of biomedical informatics*, vol. 86, pp. 25–32, 2018.

[13] C. N. Vasconcelos and B. N. Vasconcelos, “Experiments using deep learning for dermoscopy image analysis,” *Pattern Recognition Letters*, 2017.

[14] 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.

[15] A. Menegola, M. Fornaciali, R. Pires, and et al., “Towards automated melanoma screening: Exploring transfer learning schemes,” *arXiv preprint arXiv:1609.01228*, 2016.

[16] K. Shimizu, H. Iyatomi, M. E. Celebi, and et al., “Four-class classification of skin lesions with task decomposition strategy,” *IEEE Transactions on Biomedical Engineering*, vol. 62, pp. 274–283, 2015.

[17] S. Demyanov, R. Chakravorty, Z. Ge, and et al., “Tree-loss function for training neural networks on weakly-labelled datasets,” in *ISBI 2017*. IEEE, 2017, pp. 287–291.

[18] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, “Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble,” *arXiv preprint arXiv:1703.03108*, 2017.

[19] C. Barata, M. E. Celebi, and J. S. Marques, “Improving dermoscopy image classification using color constancy,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, pp. 1146–1152, 2015.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks.,” in *CVPR*, 2017, vol. 1, p. 3.

[21] N. Srivastava, G. E. Hinton, A. Krizhevsky, and et al., “Dropout: a simple way to prevent neural networks from overfitting.,” *Journal of machine learning research*, vol. 15, pp. 1929–1958, 2014.