Pattern Recognition 110 (2021) 107413

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# Explainable skin lesion diagnosis using taxonomies

Catarina Barata[a,*], M. Emre Celebi[b], Jorge S. Marques[a]

[a] Institute for Systems and Robotics, Instituto Superio Técnico, Lisboa, Portugal
[b] Department of Computer Science, University of Central Arkansas, AR, USA

ARTICLE INFO

*Article history:*
Received 15 July 2019
Revised 18 April 2020
Accepted 29 April 2020
Available online 16 May 2020

*Keywords:*
Hierarchical deep learning
Explainability
Channel attention
Spatial attention
Safety-critical CADS
Skin cancer

ABSTRACT

Deep neural networks have rapidly become an indispensable tool in many classification applications. However, the inclusion of deep learning methods in medical diagnostic systems has come at the cost of diminishing their explainability. This significantly reduces the safety of a diagnostic system, since the physician is unable to interpret and validate the output. Therefore, in this work we aim to address this major limitation and improve the explainability of a skin cancer diagnostic system. We propose to leverage two sources of information: (i) medical knowledge, in particular the taxonomic organization of skin lesions, which will be used to develop a hierarchical neural network; and (ii) recent advances in channel and spatial attention modules, which can identify interpretable features and regions in dermoscopy images. We demonstrate that the proposed approach achieves competitive results in two dermoscopy data sets (ISIC 2017 and 2018) and provides insightful information about its decisions, thus increasing the safety of the model.



## 1. Introduction

Skin cancer is one of the most common types of cancer, and one of the few whose incidence rates have been steadily increasing [1]. Thus, it is crucial to improve the diagnostic accuracy, as well as the rates of early diagnosis. Two lines of work are being pursued to address this health problem: (i) investment in newer and better imaging techniques, such as confocal microscopy and spectral imaging; and (ii) development of computer aided diagnostic systems (CADS) for the automatic analysis of dermoscopy images. In particular, the latter has seen an impressive growth in the past years [2,3], mainly due to the public release of increasingly larger data sets [4]. Another changing factor was the increase in computational power, thanks to more powerful graphical processing units (GPUs) that accelerated the development of methods based on convolutional neural networks (CNNs). These networks are able to achieve (near) human expert diagnostic performances [5,6], and are trained in an end-to-end fashion, eliminating the need for hand-crafted features [7].

The features learned by CNN models are optimal, in the sense that they are optimized to give the best classification performance. However, they are not easy to interpret, especially by non-experts, and the user is left without much information to understand the output of a CNN. In safety-critical medical applications, such as the one addressed in this paper, it is crucial for CADS to provide explainable outputs to physicians. Otherwise, an incorrect diagnosis may be rendered, incurring in high costs for both the patient and the practitioner. Our work aims to address this issue through the design of an explainable CADS.

Various approaches have been proposed by the machine learning community to improve the explainability of a CNN, most of them focused on inspecting the features learned by the model. Two popular strategies are class activation maps (CAMs [8] or Grad-CAMs [9]), which highlight the image regions that contribute the most to an output, and attention modules [10] that are trained to guide the CNN towards the most discriminative features. It is also possible to inspect each filter learned by the CNN [11,12]. Most visualization methods are applied only during the inference phase and after the network is fully trained. On the other hand, there are methods try to simultaneously improve the explainability of the CNN and its performance. In this case, the network is trained to jointly perform a set of related tasks. These multi-task networks learn better features that capture common and discriminative properties [13,14].

In this work we propose to combine multi-task CNNs with visualization methods to develop an explainable CADS for skin cancer diagnosis. Towards this goal we will take into account a property of skin lesions that remains relatively unexplored in the literature: their inherent hierarchical structure. Lesions are progressively organized by dermatologists into various classes,

* Corresponding author.
*E-mail addresses:* ana.c.fidalgo.barata@tecnico.ulisboa.pt, ana.c.fidalgo.barata@ist.utl.pt (C. Barata).

https://doi.org/10.1016/j.patcog.2020.107413
0031-3203/

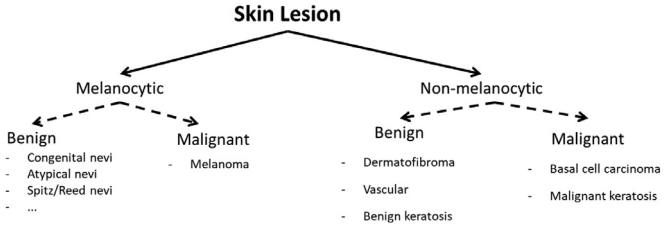


**Fig. 1.** Hierarchical organization of skin lesions.

according to their origin (melanocytic or non-melanocytic) and degree of malignancy (malignant or benign), until a differential diagnosis is reached (see Fig. 1). In order to determine these sequential classes, dermatologists screen the lesions for the presence of localized dermoscopic criteria [15]. Various dermoscopic criteria, such as streaks or blood vessels, are highly correlated with the origin of the lesion (*e.g.*, melanocytic for the streaks and non-melanocytic for blood vessels), while a more detailed assessment of the structures makes it possible for dermatologists to perform a differential diagnosis based on the following medical facts: (i) irregular streaks are a sign of melanoma, but regular ones are a hallmark of the reed and spitz nevi; (ii) arborizing vessels are associated with basal cell carcinomas, while the hairpin ones are more common in seborrheic/benign keratosis.

Expert dermatologists are able to achieve better diagnosis by understanding the aforementioned similarities and differences between the various lesions. Thus, it is expected that CADS would also benefit from this knowledge. In this work, we will develop a deep learning based CADS that makes hierarchical decisions about the lesion (multi-task) at the following levels: origin (melanocytic/non-melanocytic), degree of malignancy (benign/malignant), and differential diagnosis (*e.g.*, melanoma, basal cell carcinoma, benign keratosis), where each decision is conditioned on the previous one. To mimic the localized analysis and improve the explainability of the model, we will take advantage of attention modules. Attention will guide the model towards the most discriminative regions and features of the lesion, for each of the decision levels.

Our work demonstrates the advantages of combining a multi-task CNN with attention modules. First, we prove that an explainable hierarchical model can be efficiently trained without the need to add external data, even with a small training set (2000 images), and generalizes well to new images. The model achieves competitive diagnostic results on public data sets, especially when compared with more complex methods based on ensembles of CNNs. Second, the visualization of the attention modules allows an easy interpretation of the correct and incorrect diagnosis, increasing the safety of the model. Finally, the importance of the attention module is further supported by our robustness experiments, where we visualize the impact of various image transformations. We believe that our work is a relevant contribution towards the design of more efficient, robust, and safe deep learning models.

## 2. Related work

In recent years, the field of dermoscopy image analysis has profoundly changed by the adoption of deep learning methods. CNNs have been shown to achieve accuracies very similar to those of dermatologists on the diagnosis of multiple types of skin lesions [5,6], while in prior works the focus was mainly on the differentiation between melanoma and nevi. These studies demonstrate the ability of CNNs to learn discriminative lesion representations.

The process by which a convolutional neural learns representations and associates them with a class label is not transparent. Thus, visualization strategies like CAMs [8,9] have been proposed to identify the regions in an image that contribute to the decision. This is particularly important in medical applications, where the ability of a system to be self-explainable is becoming more and more relevant. CAMs have already been adopted in dermoscopy image analysis [16,17]. An alternative to CAMs, is to visualize the filters learned by DCNNs [11,12]. Such analysis was conducted by Van Simoens and Dhoedt for a skin cancer model [18]. Their results showed that a CNN was able to learn filters that were sensitive to: border, lesion and skin color, hair, and artifacts. However, visualization approaches only allow the inspection of the network during the inference phase, and although they improve explainability, they do not impact the performance of the network.

An alternative to the previous approaches is to design multi-task architectures to explicitly influence the performance of the CNN and achieve more interpretable models. In this case, CNNs are designed to perform more than one related task [13,14,19]. This formulation makes it possible to extract more discriminative features by incorporating complementary knowledge into the CNN, while at the same time improving the overall performance in the various tasks. This approach has also been adopted in dermatoscopy, by combining lesion diagnosis with either lesion segmentation [16] or detection of dermoscopic criteria [20,21]. However, the training of such systems may be limited by: (i) the amount of available data and its representativeness; (ii) the need for manual or semi-automatic segmentations of the lesions and/or dermoscopic criteria; and (iii) missing or noisy labels, in the case of dermoscopic criteria (*e.g.*, [20,21] solely relied on criteria associated with melanocytic lesions).

In this work, we propose to develop a multi-task CNN that performs a hierarchical diagnosis of the skin lesions (recall the taxonomy from Fig. 1). This will mimic the procedure adopted by dermatologists, where their first challenge is to identify the origin of the lesion [22], prior to any differential diagnosis. Hierarchical CNNs have been used in a multitude of fields, ranging from coarse-fine image classification [23,24] to image captioning [25,26]. The concept of hierarchy has also been addressed in a small number of dermoscopy works. Shimizu et al. [27] demonstrated that a hierarchical diagnosis using hand-crafted features and an SVM led to a better performance on the diagnosis of four types of skin lesions. Barata and Marques [28] showed that a two-level hierarchical classification could improve the performance of a CNN in the diagnosis of melanomas, nevi, and benign keratoses. However, the system lacked explainability and only one type of non-melanocytic lesions was used. Moreover, since the method is based on sequential fully connected layers, extending it to larger taxonomies would increase the model complexity significantly.

The proposed model significantly differs from the above ones. First, we will use a recurrent neural network (RNN) to perform the hierarchical diagnosis. RNNs have been shown to perform well on hierarchical classification tasks, and are easily extended to incorporate larger taxonomies [23,29]. Moreover, we address the explainability of the model by incorporating an attention module that can guide the RNN towards the most relevant regions and learn more discriminative features.

A preliminary version of this work was recently published in Barata et al. [30]. We improve our earlier work in several significant ways: (i) an extension of the attention module to comprise both spatial and channel attention, such that it is simultaneously able to identify the most relevant regions and features (channels) for each label; (ii) an extensive evaluation of the best CNN architecture for image encoding; (iii) identification of the best label inference strategy; (iv) incorporation of a hierarchical loss function based on cosine similarity; (v) inclusion of taxonomies with variable lengths; and (vi) insights on the robustness of the model w.r.t. geometrical and color transformations of the images.

## 3. Proposed system

This work proposes a new CADS for skin lesions with the following properties: (i) it mimics the hierarchical decisions made by dermatologists (recall Fig. 1), thus medical knowledge is incorporated in the design of the network; and (ii) it is explainable, since it provides visual information regarding the most relevant regions and features in each step of the diagnosis.

Hierarchical classification may be seen as a problem of finding a set of sequential class labels ($C = \{C_1, \ldots, C_T\}$) that better defines image $I$, *i.e.*, the sequence that maximizes

$$p(C|I) = \prod_{t=1}^{T} p(C_t|I, C_0, \cdots, C_{t-1}). \quad (1)$$

This formulation enforces that the $t$th class is conditioned on the previous ones, thus ensuring a hierarchical dependence between classes. The probability defined in (1) is used in various problems, such as the one of image captioning, where the goal is to predict the best sequence of words to describe an image. In this work, we take inspiration from some of the approaches to model (1). In particular, we will focus on those approaches that use deep learning models [31]. Most of these works use an RNN to capture the conditional relationship between words, where the hidden state $h_t \in \mathbb{R}^P$ of the networks is responsible for propagating its "memory".

The architecture of the proposed CADS, shown in Fig. 2, is similar to that of state-of-the-art image captioning methods [32]. The first block of the system is the image encoding one, where features are extracted from the entire images using a CNN. The last block is an RNN with a long-short term memory (LSTM) cell that is responsible for generating the sequential labels for a given dermoscopy image. At consesuctive time-steps, the LSTM takes as input image features and the labels generated in the previous steps, to predict the next one. The block in the middle, called the attention module, interacts with the LSTM to define which features will be fed to the latter, taking into account the previously generated labels. The goal of the attention module is to mimick the way dermatologists analyze dermoscopy images: selectively focus on parts of the image, according to the stage of their decision process, as defined in Fig. 1. This dynamic feature extraction is called spatial attention and allows us to obtain richer descriptions for images, while at the same time improving the explainability of the model, since the attention maps may be visualized [33]. However, solely relying on spatial attention does not allow us to take full advantage of the features extracted by the CNN [34]. As shown by Van Simoens and Dhoedt [18], the various filters of a CNN capture different properties of skin lesions, such as its border transitions or colors, and the presence of acquisition artifacts. Some of these features may be more relevant than others for the hierarchical diagnosis. Therefore, we incorporate a channel attention process in the attention module, which consists of dynamically selecting the most informative channels (activation maps) at each time-step of the LSTM. In the following section we explain each of these blocks in detail.

## 4. Hierarchical diagnosis model

The proposed hierarchical diagnosis model is formed by three main blocks, as shown in Fig. 2: (i) image encoder, which extracts image features; (ii) image decoder, which performs the hierarchical classification; and iii) attention module that guides the model towards the most discriminative features and regions according to the previous output of the LSTM. The following subsections are organized according to these blocks.

### 4.1. Image encoder

The goal of this block is to extract discriminative features from raw images. In particular, we will compare three popular CNN architectures (VGG-16 [35], DenseNet-161 [36], and ResNet-50 [37]) and select the activation maps from their last convolutional layers. These maps are then vectorized and concatenated. Thus, for each image we obtain a set of $L$ image descriptors $\mathbf{x} = \{x_1, \ldots, x_L\}$, $x_l \in \mathbb{R}^D$, where $\sqrt{L} \times \sqrt{L}$ is the shape of the activation maps and $D$ is the depth, *i.e.*, the total number of channels. The values of these parameters depend on the architecture of the image encoder: (i) for VGG-16, $L = 324$ and $D = 512$; (ii) for ResNet-50, $L = 196$, $D = 1536$; and (iii) for DenseNet-161, $L = 81$ and $D = 2208$.

### 4.2. Image decoder - hierarchical classification

This block is responsible for sequentially diagnosing the dermoscopy images, following the medical taxonomy. At each hierarchical split $t$, the LSTM produces a finer class for the skin lesion, receiving as input the previously generated class $C_{t-1}$, the context (image features) $z_t$, and the previous hidden state $h_{t-1}$. The inference process inside the LSTM cell is formulated as follows

$$\begin{aligned}
i_t &= \sigma(\mathbf{W}_{is}\mathbf{E}C_{t-1} + \mathbf{W}_{ih}h_{t-1} + \mathbf{W}_{iz}z_t + b_i),\\
f_t &= \sigma(\mathbf{W}_{fs}\mathbf{E}C_{t-1} + \mathbf{W}_{fh}h_{t-1} + \mathbf{W}_{fz}z_t + b_f),\\
c_t &= f_t c_{t-1} + i_t \tanh(\mathbf{W}_{cs}\mathbf{E}C_{t-1} + \mathbf{W}_{ch}h_{t-1} + + \mathbf{W}_{cz}z_t + b_c),\\
o_t &= \sigma(\mathbf{W}_{os}\mathbf{E}C_{t-1} + \mathbf{W}_{oh}h_{t-1} + \mathbf{W}_{oz}z_t + b_o),\\
h_t &= o_t \tanh(c_t), \quad (2)
\end{aligned}$$

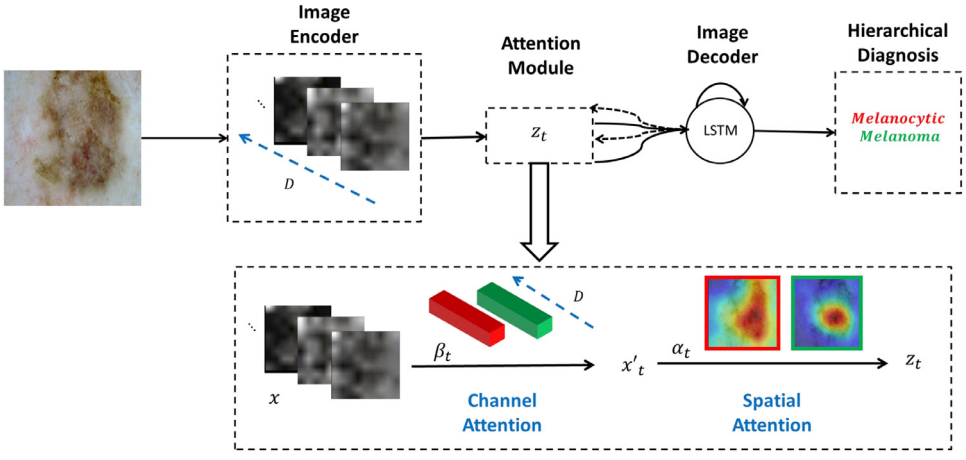


**Fig. 2.** Block diagram of the proposed system. The colors identify the channel and spatial attention weights ($\beta_t$ and $\alpha_t$) associated with each of the sequential diagnosis.

where $i_t, f_t, c_t, o_t$, and $h_t$ are the input, forget, memory, output, and hidden state of the LSTM, respectively. The network parameters (the weights $\mathbf{W}_\bullet$ and biases $b_\bullet$) are learned during the training phase, as well as the class-embedding matrix $\mathbf{E} \in \mathbb{R}^{M \times K}$, where $M$ is the class-embedding dimension and $K$ is the number of words/classes. Finally, the context vector $z_t$ is computed using the attention module, which will be discussed in the following section.

At each time step $t$, the hidden state $h_t$ is used to obtain $p(C_t|I, C_{t-1})$ and predict the next level in the taxonomy. Various strategies can be used to obtain this probability. In this work, we compare the following two:

**Class Inference - Method 1 (CI1):**

$$p(C_t|I, C_{t-1}) = \text{softmax}(\mathbf{W}_{o1} h_t), \tag{3}$$

**Class Inference - Method 2 (CI2):**

$$p(C_t|I, C_{t-1}) = \text{softmax}(\mathbf{W}_{o2}(\mathbf{W}_z z_t + \mathbf{W}_h h_t)), \tag{4}$$

where $\mathbf{W}_{o1} \in \mathbb{R}^{K \times P}$, $\mathbf{W}_z \in \mathbb{R}^{M \times D}$, $\mathbf{W}_h \in \mathbb{R}^{M \times P}$, and $\mathbf{W}_{o2} \in \mathbb{R}^{K \times M}$ are trainable weight matrices. The predicted class $\hat{C}_t$ is then selected as the one that maximizes $p(C_t|I, C_{t-1})$.

To be able to use the LSTM to predict the first level of the taxonomy, *i.e.*, at time-step $t = 1$, it is necessary to initialize its state and memory ($h_0$ and $c_0$). Following Xu et al. [33], we will use two perceptrons to infer these parameters from the average values of the activation maps obtained from the CNN: $\frac{1}{L}\sum_{l=1}^{L} x_l$.

### 4.3. Attention module

The goal of the attention module is to provide the LSTM network with the most relevant image features (called the context vector $z_t$), that can be used to predict the $t-th$ level of the taxonomy. This approach aims to mimic the analysis performed by dermatologists, while diagnosing skin lesions: inspect the lesion for localized criteria that give clues about its origin, followed by the detailed analysis of some of the identified criteria to perform the differential diagnosis. We reproduce this analysis by incorporating a spatial attention mechanism in our model. Spatial attention will enforce the LSTM network to selectively focus on different parts of the skin lesion, taking into account the previously predicted hierarchical labels. Additionally, recent works, such as that of Van Simoens and Dhoedt [18], have shown that CNN filters respond to different properties of the skin lesions. Since some of these filters may convey more discriminative information than others, we also incorporate a channel attention mechanism in our model. Similarly to the spatial attention block, the channel one is also influenced by the previously generated hierarchical labels and dynamically provides the LSTM with the most relevant image features (activation maps) to predict the following labels. Below we provide details about these two attention mechanisms.

Channel and spatial attention have been combined in various frameworks, such as those proposed in Chen et al. [34], Woo et al. [38]. Both of these works have demonstrated that the best performances are achieved when spatial attention is applied to the output of a channel attention block. Therefore, we also adopt this organization. Channel attention has been addressed in the literature as a technique to perform feature selection and to improve the representation power of the model [38]. The non-negative channel attention weights $\beta_t$ are computed as follows

$$\begin{aligned} \beta_t &= \sigma(\mathbf{W}_{ca}(\tanh(\mathbf{W}_{cax}\tilde{\mathbf{x}} + \mathbf{W}_{cax}\tilde{\mathbf{x}} + \mathbf{W}_{cah} h_{t-1}))), \\ \mathbf{x}'_t &= \beta_t \otimes \mathbf{x} \end{aligned} \tag{5}$$

where $\sigma$ is the sigmoid function, $\otimes$ denotes the element-wise multiplication, $\bar{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ are respectively the average and max-pooled feature maps, and the matrices $\mathbf{W}_{ca}$, $\mathbf{W}_{cax} \in \mathbb{R}^{D \times \frac{D}{r}}$, and $\mathbf{W}_{cah} \in$ are all trainable. The new features $\mathbf{x}'_t$ can be used to compute the spatial attention weights as described next.

Pure spatial attention methods can be divided into hard and soft attention [33]. Soft attention, as proposed by Chorowski et al. [39] is used in this work. Based on this method, we estimated the context vector $z_t$ as follows

$$z_t = \sum_{i=1}^{L} \alpha_{ti} x'_{ti}, \tag{6}$$

where $\alpha_t = \{\alpha_{t1}, \ldots, \alpha_{tL}\}$ is a set of non-negative weights that represent the relative importance of each position in the feature map.

In this work, we apply the method proposed by Xu et al. [33] to estimate the weights $\alpha_t$ as follows

$$\alpha_t = \text{softmax}(\mathbf{W}_{sa}(\tanh(\mathbf{W}_{sax}\mathbf{x}' + \mathbf{W}_{sah} h_{t-1}))), \tag{7}$$

where $\mathbf{W}_a$, $\mathbf{W}_{ax}$, and $\mathbf{W}_{ah}$ are all trained in parallel with the other parameters of the model.

The outputs of both attention mechanisms may be visually inspected, *i.e.*, we can identify which were the activation maps with the highest values in $\beta_t$ and visualize the spatial map $\alpha_t$, for each of the predicted labels. This greatly improves the explainability of the model, since we are able to provide visual cues that justify each of the hierarchical decisions.

## 5. Experimental setup

### 5.1. Data set and experiments

We developed our model using the ISIC 2017 and 2018 dermoscopy data sets [4,40]. The first set comprises 2750 images divided into training (2000), validation (150), and test (600) sets. These images contain examples of the following classes of lesions: melanocytic (melanoma and nevi) and non-melanocytic (seborrheic keratosis). The second set is larger and more complex, containing 11,527 examples of the following lesions: melanocytic (melanoma and nevi) and non-melanocytic (basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, and vascular lesions). Similarly to ISIC 2017, the 2018 data set is also divided into training (10,015) and test sets (1512). Neither data set was augmented with external data and, in order to deal with image color variability induced by different acquisition setups, all of the images were normalized using the color normalization scheme proposed in Barata et al. [41]. Fig. 3 shows some examples of normalized images.

We report the results of several ablation studies conducted on the ISIC 2017 data set, namely:

i) **Image encoder:** The representation power of the three CNN architectures (DenseNet-161, VGG-16, and ResNet-50) was evaluated by using each network separately as an image encoder.

ii) **Hierarchical loss function:** A commonly used loss function in captioning problems is the categorical cross-entropy. However, by using this loss function alone there is no guarantee that the model will be able to learn the taxonomic structure of the various classes *i.e*, the sequential path between the coarse and finer classes. Thus, we propose to explicitly model the taxonomic constraints on the loss function and compare this approach against solely using the cross-entropy loss. Details about the hierarchical loss function will be given in the following subsection.

iii) **Channel attention:** The combined use of the channel and spatial attention mechanisms (recall Section 4.3) was evaluated for several values of the channel-reduction factor $r \in \{1, 2, 4, 8, 16\}$, where $\frac{1}{r}$ denotes the proportion of channels that the model should select. Setting $r = 1$ is equivalent to using only spatial attention.

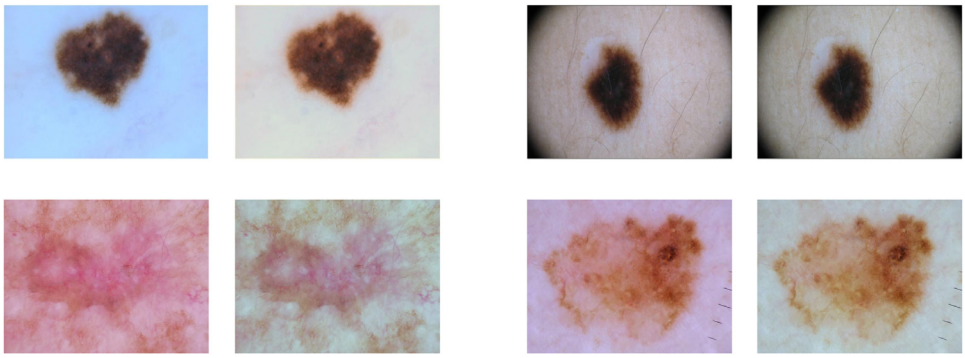

**Fig. 3.** Examples of dermoscopy images (1st column) and their color normalized versions using: [41] (2nd column): ISIC 2017 (1st row) and ISIC 2018 (2nd row).

iv) **Class inference:** The two class inference approaches presented in Section 4.2 are compared to determine if the predicted state $h_t$ conveys sufficient information for the diagnosis or if it is beneficial to add context $z_t$ to the classification model.

v) **Model hyperparemeters:** The impact of various model hyperparameters in the performace of the hiearchical diagnostic system are assessed. In particular, we compare the performance for different values of $M$ (size of the lesion class-embedding) and $P$ (size of the hidden state $h$ of LSTM).

We train a hierarchical model for each of the aforementioned configurations, using the ISIC 2017 training set. The validation set is used for early stopping and selection of the best hyperparameters, while the test set is used to quantify the performance of the model. We would like to emphasize that we do not use any external data set to augment the ISIC 2017 training set.

The ISIC 2018 data set was used to assess the performance of the model in a more complex scenario, where there are more types of non-melanocytic lesions (this data set contains examples of five types of non-melanocytic lesions, while ISIC 2017 contained only one). Additionally, we conduct the following experiments:

**a) Varying taxonomy length:** The ISIC 2018 data set contains seven classes of skin lesions, five of which are non-melanocytic. Among the non-melanocytic lesions, one can further split them into benign (keratosis, dermatofibroma, and vascular) and malignant (BCC and actinic) lesions, prior to their differential diagnosis. In order to infer if this additional hierarchical class helps the model to learn more discriminative representations of the skin lesions, we will compare the following approaches: (i) two level taxonomy ($T = 2$), where two hierarchical decisions are performed (melanocytic/non melanocytic and the differential diagnosis); and (ii) varying length, where the model has first to classify the lesion as melanocytic/non melanocytic and then, if the lesion is classified as melanocytic, perform a differential diagnosis (melanoma/nevu), otherwise it should first classify the lesion as malignant/benign and only then the differential diagnosis.

**b) Changing the taxonomic order:** The taxonomy defined in Fig. 1 has been defined by dermatologists to take into account the origin of the lesions as the first splitting criterion. However, often the decision to excise the lesion is based on whether or not it shows signs of malignancy. Only at the pathologist level will the lesion be diagnosed according to its origin and differential class. To mimic this process and determine its influence on the performance of the model, we change the order of the hierarchical decisions, *i.e.*, first classify the lesion and malignant/benign, then diagnose it as melanocytic or non-melanocytic, and finally perform the differential diagnosis.

**c) Robustness experiments:** Recently, various authors have reported improved performances by augmenting the test set with geometric transformations, such as central crops of different ratios and horizontal/vertical flips (*e.g.* [42]), followed by an ensemble like classification of the augmented images. Although the results are promising, it also suggests that the learned models may not be robust to changes in viewpoint. This is undesirable, since it means that the same lesion may be assigned a different label depending on its acquisition. Thus, we evaluate the robustness of our method w.r.t to: i) performing a horizontal flip on the images; ii) performing a central crop to get only 80% of the original image; and iii) removing the color normalization exemplified in Fig. 3, *i.e.*, applying the model directly to the original image.

Unfortunately, the ground truth information (i.e., diagnostic labels) for the test set of ISIC 2018 is not publicly available. The performance on this set can only be assessed through an online platform. To be able to conduct the same training procedure as in the case of ISIC 2017, we used 5-fold cross validation on the training set. Then, each of the five models was used to classify the images on the test set and the final diagnosis was determined by averaging the scores of the models.

### 5.2. Model training

Each of the model's configurations discussed in the previous section was trained end-to-end using a data set $\mathcal{I} = \{I^1, \ldots, I^N\}$, for which we have the corresponding ground truth sequences of hierarchical labels $\mathcal{C} = \{C^1, \ldots, C^N\}$. In order to train the model, we apply the balanced categorical cross-entropy loss to each training example [43]

$$\mathcal{L}_{CE}(I, C) = -\frac{1}{T}\sum_{i=1}^{T} w_{C_t} \log p(C_t | I, C_{t-1}), \tag{8}$$

where $T$ is the length of the taxonomy of the example, $w_{C_t}$ is the label-weight, and $p(C_t|I, C_{t-1})$ is given by either (3) or (4). In our models, we set $T = 2$ or 3, *i.e.*, we either consider two levels in the taxonomy (melanocytic/non-melanocytic and the differential diagnosis) or two levels for the melanocytic lesions and three for the non-melanocytic ones. The label-weights $w_{C_t}$ are used to deal with the severe imbalance of classes in the training set, and are given by

$$w_{C_t} = \frac{\#N}{\#N_c}, \tag{9}$$

where #N is the size of the training set and $\#N_c$ is the number of examples in each class. To ensure that these weights do not influence the learning rate of the stochastic gradient descent, we normalize the weights in each batch such that they have unit mean.

Despite being used in classification and captioning problems, solely relying on the cross-entropy loss may be sub-optimal in our hierarchical problem, since this loss only penalizes incorrect labels and not deviations from the taxonomic structure (*e.g.* reaching the

**Table 1**
Comparision of loss functions - best scores on the ISIC 2017 test set using CI1 as inference approach. HL stands for hierarchical loss.

| Encoder | Lesion class | HL | *SE* | *SP* | *BACC* | *AUC* |
|---|---|---|---|---|---|---|
| VGG-16 | **MEL/NON-MEL** (#510/#90) | | 87.2% | 76.7% | 82.0% | 92.4% |
| | Keratosis (#90) | | 75.5% | 86.9% | 70.7% | 91.6% |
| | Melanoma (#117) | | 63.2% | 85.5% | | 78.8% |
| | Nevus (#393) | | 73.3% | 82.6% | | 83.7% |
| | **MEL/NON-MEL** (#510/#90) | ✓ | 86.9% | 74.4% | 80.6% | 91.6% |
| | Keratosis (#90) | ✓ | 74.4% | 86.9% | 68.8% | 90.6% |
| | Melanoma (#117) | ✓ | 53.8% | 89.2% | | 78.0% |
| | Nevus (#393) | ✓ | 82.6% | 78.1% | | 84.0% |
| ResNet-50 | **MEL/NON-MEL** (#510/#90) | | 82.1% | 80.0% | 81.1% | 90.2% |
| | Keratosis (#90) | | 80.0% | 82.3% | 67.1% | 90.3% |
| | Melanoma (#117) | | 48.7% | 87.8% | | 75.8% |
| | Nevus (#393) | | 72.5% | 82.1% | | 84.3% |
| | **MEL/NON-MEL** (#510/#90) | ✓ | 83.1% | 80.0% | 81.6% | 90.3% |
| | Keratosis (#90) | ✓ | 80.0% | 83.1% | 67.8% | 90.8% |
| | Melanoma (#117) | ✓ | 50.4% | 89.0% | | 77.2% |
| | Nevus (#393) | ✓ | 73.0% | 79.2% | | 83.6% |
| DenseNet-161 | **MEL/NON-MEL** (#510/#90) | | 90.8% | 70.0% | 80.4% | 90.4% |
| | Keratosis (#90) | | 68.9% | 90.8% | 68.5% | 90.9% |
| | Melanoma (#117) | | 54.7% | 88.6% | | 82.4% |
| | Nevus (#393) | | 81.9% | 75.8% | | 87.0% |
| | **MEL/NON-MEL** (#510/#90) | ✓ | 88.6% | 84.4% | 86.5% | 93.6% |
| | Keratosis (#90) | ✓ | 84.4% | 88.6% | 73.0% | 93.3% |
| | Melanoma (#117) | ✓ | 55.6% | 89.6% | | 81.7% |
| | Nevus (#393) | ✓ | 79.1% | 80.2% | | 86.3% |

correct differential diagnosis through an incorrect path). To address this issue, we also propose to use a hierarchical loss function, defined as follows

$$\mathcal{L}(I,C) = \mathcal{L}_{CE}(I,C) + \mathcal{L}_H(I,C), \tag{10}$$

where $\mathcal{L}_H(I,C)$ is given by the cosine distance

$$\mathcal{L}_H(I,C) = 1 - \frac{H \cdot \hat{H}}{\|H\|\|\hat{H}\|}. \tag{11}$$

Here $H \in \mathbb{R}^A$ is a binary path-vector, $A$ is the set of labels in the taxonomy tree, and $H_a = 1$ if the $a-th$ label belongs to the taxonomic classification of the lesion. Finally, $\hat{H} \in \mathbb{R}^A$ is the estimated path.

The model is optimized end-to-end using the Adam variation of the stochastic gradient descent [44] with mini-batches of size 20, using an initial learning rate of $10^{-6}$, which decays at every 200 epochs. In total, the model is trained for 600 epochs with an early-stop criterion on a NVIDIA Titan Xp[1]. The model parameters are set as follows: $M \in \{50, 100, 300\}$, and $P$ tuned in the interval $\{2^8, \ldots, 2^{10}\}$. To improve the generalization of the model, we have adopted the following strategies: (i) careful initialization of several model weights $\mathbf{W}_\bullet$; (ii) online data augmentation (a sequence of random crop, random flip, and random color transformation at each epoch); and iii) incorporation of dropout with 50% probability in several of the layers. The carefully initialized weights are those of the image encoder (CNNs), where we used the weights of the models pre-trained on ImageNet, and those of the word encoding $\mathbf{E}$ that were initialized from the GloVe embeddings [45].

### 5.3. Model evaluation

All of the trained model configurations were evaluated using an independent test set, and the performance was quantified using the following metrics: sensitivity (*SE*), specificity (*SP*), balanced accuracy (*BACC*), and area under the curve (AUC). The metrics *SE*, *SP*, and *AUC* are class specific, while *BACC* is computed over the entire data set.

For the ISIC 2017 data set, we also compare our results with others recently reported in the literature. In particular, we establish comparisons with the results of: (i) Harangi [46], who compared multiple CNN architectures using the ISIC 2017 data set but does not use a hierarchy; (ii) Barata and Marques [28], who investigated the inclusion of a class-hierarchy in a CNN framework using two levels of fully connected layers; and (iii) the top ranked participants of the ISIC 2017 challenge.

## 6. Experimental results

### 6.1. Ablation studies on ISIC 2017

In this section, we report the results for the several ablation studies described in Section 5.1. We will report the results of the studies as follows. First, we show the results for the analysis of the loss function for all image encoders, but using only CI1 (3) as the class inference method, since the performance for CI2 (4) was similar. We then evaluate the performance after the incorporation of the channel attention module, using all of the image encoders. We also use this scenario to compare the inference strategies CI1 and CI2. Finally, we select the two best models to perform a detailed analysis of the influence of the model parameters: $P$ (the size of the LSTM hidden layer) and $M$ (the size of the lesion class embedding).

**Hierarchical loss function:** Table 1 shows the scores for models trained using either the cross-entropy loss (8) or the hierarchical loss (10).

By inspecting Table 1 we can observe that enforcing the hierarchical structure in the loss function improves the performance of the models, particularly when ResNet-50 and DenseNet-161 are used as image encoders (*e.g.*, note the improvement in the *BACC* scores). Moreover, if we compare the *SE* and *SP* values of the keratosis class (the only non-melanocytic lesion) with those of the melanocytic/non-melanocytic task, we conclude that only the hierarchical loss (10) guarantees that the model does not violate the taxonomy.

[1] The source code will be available on https://github.com/catarina-barata/skin-hierarchy/

**Table 2**
Comparison of channel attention and inference strategies (termed configuration) - best scores on the ISIC 2017 test set.

| Encoder | Lesion class | Config. | *SE* | *SP* | *BACC* | *AUC* |
|---|---|---|---|---|---|---|
| VGG-16 | **MEL/NON-MEL** (#510/#90) | | 87.1% | 86.7% | 86.9% | 92.4% |
| | Keratosis (#90) | CI1, $r = 4$ | 86.7% | 87.1% | 74.3% | 91.6% |
| | Melanoma (#117) | | 60.7% | 89.7% | | 80.0% |
| | Nevus (#393) | | 75.6% | 83.1% | | 84.7% |
| | **MEL/NON-MEL** (#510/#90) | | 86.7% | 77.8% | 82.2% | 91.2% |
| | Keratosis (#90) | CI2, $r = 2$ | 76.7% | 86.8% | 70.7% | 91.0% |
| | Melanoma (#117) | | 58.1% | 90.3% | | 83.3% |
| | Nevus (#393) | | 77.3% | 78.3% | | 85.9% |
| ResNet-50 | **MEL/NON-MEL** (#510/#90) | | 85.3% | 80.0% | 82.6% | 91.5% |
| | Keratosis (#90) | CI1, $r = 2$ | 80.0% | 85.3% | 69.1% | 91.3% |
| | Melanoma (#117) | | 52.1% | 88.0% | | 77.5% |
| | Nevus (#393) | | 75.3% | 81.6% | | 85.9% |
| | **MEL/NON-MEL** (#510/#90) | | 88.8% | 78.9% | 83.9% | 91.3% |
| | Keratosis (#90) | CI2, $r = 2$ | 80.0% | 88.8% | 70.8% | 90.9% |
| | Melanoma (#117) | | 53.0% | 90.5% | | 80.2% |
| | Nevus (#393) | | 79.4% | 75.4% | | 85.0% |
| DenseNet-161 | **MEL/NON-MEL** (#510/#90) | | 89.4% | 82.2% | 85.8% | 94.1% |
| | Keratosis (#90) | CI1, $r = 2$ | 82.2% | 89.6% | 73.4% | 93.5% |
| | Melanoma (#117) | | 59.0% | 88.6% | | 80.0% |
| | Nevus (#393) | | 79.1% | 81.6% | | 86.1% |
| | **MEL/NON-MEL** (#510/#90) | | 93.3% | 70.0% | 81.8% | 92.1% |
| | Keratosis (#90) | CI2, $r = 4$ | 68.9% | 92.9% | 72.2% | 91.7% |
| | Melanoma (#117) | | 70.9% | 85.1% | | 85.0% |
| | Nevus (#393) | | 76.8% | 78.3% | | 85.9% |

According to these results, it seems that ResNet-50 performs worse than the remaining image encoders and that DenseNet-161 seems to be the encoder that better captures the properties of the various classes.

**Channel attention and inference approaches:** Table 2 shows the best performances across all models trained with various degrees of channel reduction $r$ and inference strategies.

First, let us compare the results of Table 2 w.r.t to CI1, against those reported in Table 1 for the hierarchical loss. Incorporating channel attention in the model seems to improve the performance of all configurations (see *BACC* scores and the *SE* for the various classes). In particular, channel attention significantly improves the performance of the VGG-16 model. This suggests that some of the channels contain redundant information and can be discarded. Interestingly, the best ratio $r$ seems to be either 2 or 4 and larger values led to a significant performance degradation. This may be due to the variability across dermoscopy images or to the size of the training set, which does not allow the training of larger channel-attention modules.

Between the two inference strategies, CI2 seems to improve the classification scores for the melanoma class. However, the performance for the remaining classes degrades and some of the decisions violate the taxonomy (*e.g.*, compare the *SP* of the non-melanocytic lesions with the *SE* of the keratosis class). Thus, we conclude that it is preferable to use only the state of the LSTM to infer the class, *i.e.*, inference method CI1 (3). The remaining results will be reported for this formulation.

**Model hyperparameters:** Table 3 shows a comparison of the *BACC* scores for different values of $P$ and $W$. For simplicity, we limit our analysis to the two best performing image encoders: VGG-16 and DenseNet-161. Both networks seem to achieve better performances when the size of the hidden state of the LSTM is $P \geq 512$, thus we opt to use $P = 512$ to reduce the number of parameters. Regarding the size of the class-embedding $M$, the two models show distinct behaviors. The performance of the model that uses DenseNet-161 degrades, possibly due to the smaller size of the activation maps that are obtained with this network ($9 \times 9$ *vs.* $18 \times 18$ for VGG-16) and significantly higher number of filters ($D = 2208$ *vs.* $D = 512$ for VGG-16).

**Table 3**
Influence of model hyperparameters: size of the hidden state $P$ and class-embedding $M$.

| Encoder | *M* | *P* | *BACC* |
|---|---|---|---|
| VGG-16 | 50 | 256 | 72.6% |
| | 50 | 512 | 73.0% |
| | 50 | 1024 | 73.3% |
| | **100** | **512** | **74.3%** |
| | 300 | 512 | 72.2% |
| DenseNet-161 | 50 | 256 | 70.3% |
| | **50** | **512** | **73.4%** |
| | 50 | 1024 | 73.3% |
| | 100 | 512 | 71.5% |
| | 300 | 512 | 70.9% |

**Bold** highlights the best results.

### 6.2. Qualitative assessment of attention modules

In this section, we visualize the channel and spatial attention maps obtained using the VGG-16 and the DenseNet-161 based models. For the sake of simplicity, we will not discuss the model based on ResNet-50, since this was the encoder that achieved the worst overall performance. The discussed examples are all from the ISIC 2017 data set.

Figs. 4 and 5 show the spatial attention maps obtained with VGG-16 and DenseNet-161 as image encoders. Two observations can be made regarding the attention maps: i) VGG-16 (top maps) leads to more detailed maps, although it achieves slightly worse diagnostic performances; and ii) both models "attend" to regions that are relevant for the medical diagnosis and are able to show regions of interest, even when the lesion is difficult to distinguish from the surrounding skin (see the keratosis example on Fig. 4). To understand the importance of the regions identified by the models, we can take a closer look at the melanocytic images. In both examples, the models "attend" to areas that contain a dermoscopic structure called pigment network, which is considered to be one of the hallmarks of melanocytic lesions [15]. As its name suggests, pigment network consists of a mesh of dark lines over a lighter background. In Fig. 4, the pigment network areas are lo-

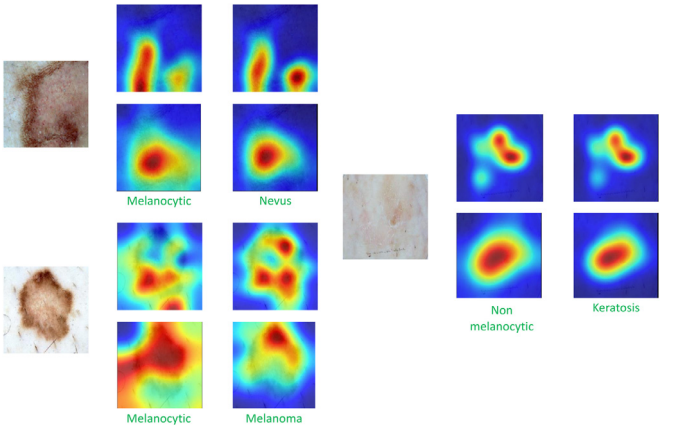


**Fig. 4.** Correctly diagnosed skin lesions from ISIC 2017 test set and the corresponding spatial attention maps for the following image encoders: VGG-16 (top maps) and DenseNet-161 (bottom maps).

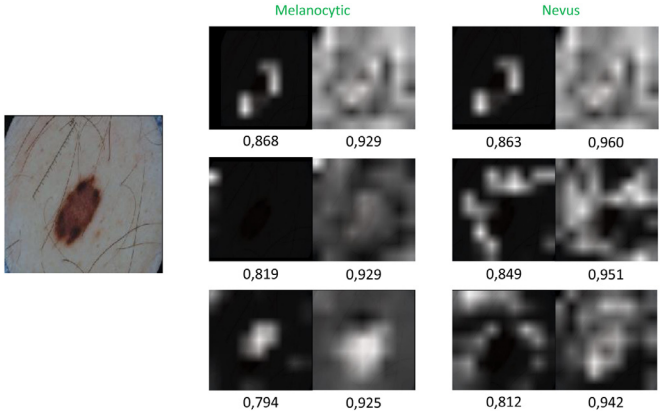


**Fig. 6.** Highest probability activation maps (channels) for a correctly diagnosed nevus. The maps for VGG-16 and DenseNet-161 and their corresponding probabilities are shown on the left and right columns, respectively.

cated near the border of the two melanocytic lesions. Pigment network may also be used to diagnose skin lesions, as a network with an atypical shape (enlarged and irregular lines) is taken as a sign of melanoma. This may justify why a benign lesion was diagnosed as melanoma in Fig. 5 (left), since both networks select one of the regions with atypical network, to perform the differential diagnosis.

Spatial attention may also be very useful to help us identify bias in the data set, as is exemplified in Fig. 5 (right). Several keratosis images from the ISIC 2017 data set contain illumination artifacts that were detected by the attention modules, allowing the model to learn incorrect features from the data.

Fig. 6 shows a comparison between the three channels with the highest weights $\beta_t$ (computed using (5)), for VGG-16 and DenseNet-161. To improve the visualization, we expanded the channel activation maps to the size of the original dermoscopy image, The maps from VGG-16 seem to capture more localized information than those from DenseNet-161. This may be due to the properties of the two networks, since in DenseNet, the maps from top layers are propagated to the deeper ones. Nonetheless, in both cases it is possible to see that the networks extract information related with the surrounding skin, the border of the lesion and its center, and the presence of hair. These findings confirm those reported by Van Simoens and Dhoedt [18], who observed that the convolutional layers were sensitive to several visual cues, including the aforementioned ones.

### 6.3. Comparison with other works

Table 4 compares our best performing ISIC 2017 models to various state-of-the-art methods. ISIC 2017 data set was released as part of a challenge, thus most of the works in the literature only report performance scores for the melanoma and keratosis classes, which were the metrics used to rank the participants. For comparison purposes, we also adopt those metrics. Additionally, we report two criteria that may influence the performance of the models: (i) ensembles of networks; and (ii) augmenting the training set with external data.

By comparing the scores of our hierarchical models with flat models that use the same CNN architecture and do not use external data [28,46], we are able to see that imposing a hierarchical diagnosis leads to impressive improvements in the performance of the model, since *SE* and *SP* scores for both classes are significantly higher for all of our models. When external data is used for training (see [17]), the flat model performs better than our hierarchical approach (see the results with ResNet-50). Nonetheless, our approach is still able to achieve a competitive performance, using less training data. Once again, this suggests that imposing a hierarchical decision process allows the training of more generalizable models.

The methods that ranked in the first four positions of the ISIC 2017 challenge used external data to train the models. Moreover, three of these works also adopted an ensemble strategy to predict the final diagnosis. When we compare our score with theirs, it is possible to see that only the top ranked method achieved better *SE* scores. Ensembles may be used to improve the representation power of a model by combining the capabilities of different CNN architectures. Our results suggest that imposing a hierarchical diagnosis allows a single architecture to learn better representations of the various lesion classes.

A final comparison can be made between the proposed models and the hierarchical approach of [28], which sequentially adds

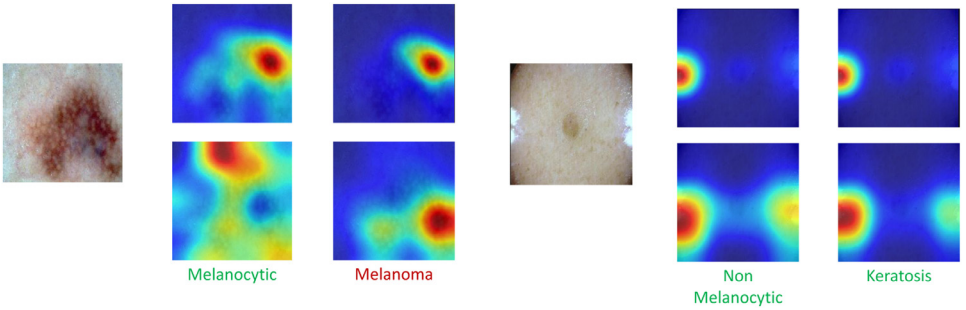


**Fig. 5.** Incorrectly diagnosed skin lesion (left) and data set bias (rights), and their corresponding spatial attention maps for the following image encoders: VGG-16 (top maps) and DenseNet-161 (bottom maps).

**Table 4**
Comparison with other works on the ISIC 2017 test set. The table is organized by CNN architecture and the methods identified with numbers correspond to the methods that ranked in the first four positions in the ISIC 2017 challenge.

| | | | Melanoma | | | Keratosis | | |
|---|---|---|---|---|---|---|---|---|
| CNN/Work | Ens. | Ext. Data | *SE* | *SP* | *AUC* | *SE* | *SP* | *AUC* |
| VGG-16 [46] | N | N | 25.6% | 58.5% | 76.6% | 61.1% | 68.6% | 82.5% |
| VGG-16 Ours | N | N | 60.7% | 89.7% | 80.0% | 86.7% | 87.1% | 92.4% |
| ResNet-50 [46] | N | N | 38.5% | 43.7% | 75.7% | 65.6% | 75.9% | 86.1% |
| ResNet-50 [17] | N | Y | 65.8% | 89.6% | 87.5% | 87.8% | 86.7% | 95.8% |
| ResNet-50 Ours | N | N | 52.1% | 88.0% | 77.5% | 80.0% | 85.3% | 91.3% |
| DenseNet-161 [28] | N | N | 46.1% | 92.5% | 81.8% | 77.8% | 90.5% | 92.1% |
| Hier. DenseNet-161 [28] | N | N | 35.9% | 94.2% | 80.0% | 71.1% | 90.9% | 90.8% |
| DenseNet-161 Ours | N | N | 59.0% | 88.6% | 80.0% | 82.2% | 89.6% | 93.5% |
| #1 [40] | Y | Y | 73.5% | 85.1% | 86.8% | 97.8% | 77.3% | 95.3% |
| #2 [40] | N | Y | 10.3% | 99.8% | 85.6% | 17.8% | 99.8% | 96.5% |
| #3 [40] | Y | Y | 54.7% | 95.0% | 87.4% | 35.6% | 99.0% | 94.3% |
| #4 [40] | Y | Y | 42.7% | 96.3% | 87.0% | 58.9% | 97.6% | 92.1% |

**Table 5**
Classification scores on the ISIC 2018 test set using different taxonomies: a) $T = 2$ two-level taxonomy (melanocytic/non-melanocytic and differential diagnosis); b) $T = 3$ three-level taxonomy (melanocytic/non-melanocytic and benign/malignant for non-melanocytic lesions, prior to the differential diagnosis); c) *Inv.* inverted taxonomy (malignant/benign followed by melanocytic/non-melanocytic and differential diagnosis).

| | $T = 2$ | | | | $T = 3$ | | | | *Inv* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lesion class | *SE* | *SP* | *BACC* | *AUC* | *SE* | *SP* | *BACC* | *AUC* | *SE* | *SP* | *BACC* | *AUC* |
| Melanoma | 57.3% | 93.1% | 72.3% | 87.7% | 67.8% | 90.1% | 72.6% | 86.1% | 59.1% | 94.6% | 71.7% | 88.9% |
| Nevus | 83.9% | 93.0% | | 96.2% | 82.0% | 93.7% | | 96.1% | 89.4% | 89.7% | | 96.0% |
| BCC | 75.3% | 98.0% | | 97.7% | 74.2% | 98.4% | | 98.5% | 76.3% | 98.4% | | 98.3% |
| Actinic | 55.8% | 99.2% | | 96.3% | 60.5% | 99.5% | | 95.5% | 62.8% | 99.0% | | 95.6% |
| Keratosis | 76.0% | 95.8% | | 95.0% | 72.8% | 96.6% | | 94.4% | 76.0% | 95.8% | | 94.6% |
| Derm. | 63.6% | 99.7% | | 98.1% | 68.2% | 99.8% | | 97.2% | 68.2% | 99.8% | | 96.7% |
| Vascular | 60.0% | 99.7% | | 97.7% | 57.1% | 99.7% | | 98.5% | 62.9% | 99.8% | | 97.8% |
| Average | 67.4% | 96.9% | | 97.7% | 69.0% | 96.8% | | 95.2% | 70.7% | 96.7% | | 95.4% |

fully-connected layers to make hierarchical decisions. The proposed approach outperforms the one presented in Barata and Marques [28], suggesting that an LSTM is more suitable for hierarchical classification.

### 6.4. ISIC 2018 - more classes and robustness

Table 5 shows the best performances on the ISIC 2018 test set, using different taxonomic lengths and orders. These results were obtained with the following configuration (selected by cross-validation): DenseNet-161 as the image encoder, channel ($r = 2$) and spatial attention, and CI1 for class inference. Since the ground truth diagnosis is not publicly available, it is not possible to evaluate the model on the other hierarchical decisions for the test set.

Similarly to what was observed for the ISIC 2017 data set, melanoma remains one of the most challenging classes. This was expected, as this type of skin cancer often mimics other types of lesions [15]. The other difficult classes are actinic, dermatofibroma, and vascular, while the model is able to achieve good performances for nevus, keratosis, and BCC. Actinic, dermatofibroma, and vascular lesions correspond respectively to 3.3%, 1.2%, and 1.4% of the training set, which is an extremely imbalanced scenario. Nevertheless, although we do not conduct any form of data augmentation, our approach is still able to achieve sensitivities above 50% for each of the minority classes.

It is possible to appreciate the influence of a hierarchical structure in the performance of the system by comparing several taxonomic configurations. Just by increasing the taxonomic level from $T = 2$ to $T = 3$ we are able to significantly improve the diagnostic performance for two of the malignant classes (melanoma and actinic), while the performance for BCC remains almost the same. There was also a slight improvement on the *BACC* score, as well as on the average *SE* for all the classes. If we modify the taxonomic order, it is possible to improve the performance for almost all lesion types, when compared with $T = 2$. However, this configuration leads to poorest melanoma scores than $T = 3$, which is undesirable as melanoma is a very aggressive form of cancer, as well as worse *BACC*. Analyzing the channel and spatial attention maps of the lesions that were incorrectly classified could allow us to further understand the performance differences among for the various taxonomic configurations, but unfortunately we do not have access to the ground truth labels for the images on the ISIC 2018 test set.

Similarly to ISIC 2017, in this case it is also possible to compare our results with those reported in the literature. The *BACC* scores for the ISIC 2018 test set fall in the range [13.2%, 88.5%],[2] which clearly shows how challenging this problem is. As in the case of the ISIC 2017 challenge, several of the best performing methods are large ensemble models that were trained with augmented versions of the original training set, which is not the case for the proposed approach.

In order to evaluate the robustness of the method to viewpoint and acquisition changes, we applied the model with $T = 3$ to transformed versions of the test set, as described in Section 5.1. The results are shown in Table 6. The sets with geometric transformations lead to slight improvements over the *BACC* score for the original set, but the performances remain fairly similar for all lesion types. On the other hand, removing the color normalization has a significant effect on the performance of the model: the *SE* for keratosis and vascular lesions drop drastically, while the *SE* for melanoma increases. Although we have conducted a form of data augmentation that used random color transformations to improve the robustness of the network, this result suggests that the fea-

[2] Source: https://challenge2018.isic-archive.com/

**Table 6**
Classification scores on the modified ISIC 2018 test sets.

| | Flip | | | 80% Central crop | | | Non-normalized images | | |
|---|---|---|---|---|---|---|---|---|---|
| Lesion class | *SE* | *SP* | *BACC* | *SE* | *SP* | *BACC* | *SE* | *SP* | *BACC* |
| Melanoma | 67.8% | 88.7% | 73.2% | 60.2% | 93.0% | 73.1% | 83.0% | 76.9% | 66.6% |
| Nevus | 80.2% | 95.4% | | 84.5% | 91.7% | | 74.1% | 93.9% | |
| BCC | 75.3% | 98.4% | | 77.4% | 98.0% | | 76.4% | 98.2% | |
| Actinic | 58.1% | 99.6% | | 69.8% | 98.6% | | 62.8% | 99.1% | |
| Keratosis | 74.2% | 95.3% | | 67.7% | 97.3% | | 33.6% | 99.5% | |
| Derm. | 70.5% | 99.7% | | 65.9% | 99.8% | | 56.8% | 99.9% | |
| Vascular | 65.7% | 99.8% | | 60.0% | 99.7% | | 48.6% | 99.7% | |

tures extracted by CNN are color sensitive. These results confirm the findings of other works in the literature, such as [47], who found out that color normalization has a positive impact on the performance of CNN. Two relevant points come out of this experiment. First, it is important to standardize the color of dermoscopy images through normalization. Second, future work should address the perception of color by CNNs, in order to make them more robust to color changes. A few works investigated the sensitiveness of CNNs to color using real images. However, to the best of our knowledge such a study is still missing in the dermoscopy field. Based on our experimental findings, we believe that this is a necessary direction, in order to improve the safety of the method.

The explainability of our model can be used to understand the aforementioned performance differences. Thus, we selected two examples from one of the cross-validation data sets: one that is consistently diagnosed as melanoma and another that is diagnosed as BCC for the original and flipped images, and as vascular for the remaining transformations. For each of the examples, we inspected their spatial attention maps as well as the most probable activation map given by channel attention. Fig. 7 shows the results, for simplicity we only show the maps for two of the decisions: melanocytic/non-melanocytic and the differential diagnosis. In all of the scenarios, the decision between melanocytic/non-melanocytic is made taking into account the skin that surrounds the lesions, as shown in the corresponding spatial attention maps. The spatial maps for the differential diagnosis are consistent across transformations, *i.e.*, on the left example the model consistently "attends" to the darkest regions, while on the right example it focuses on the blue-oval area. The main difference lies in the selected channels. These maps show activations significantly different from those obtained for the original image (1st row), demonstrating the lack of robustness of the CNN architecture to these transformations.

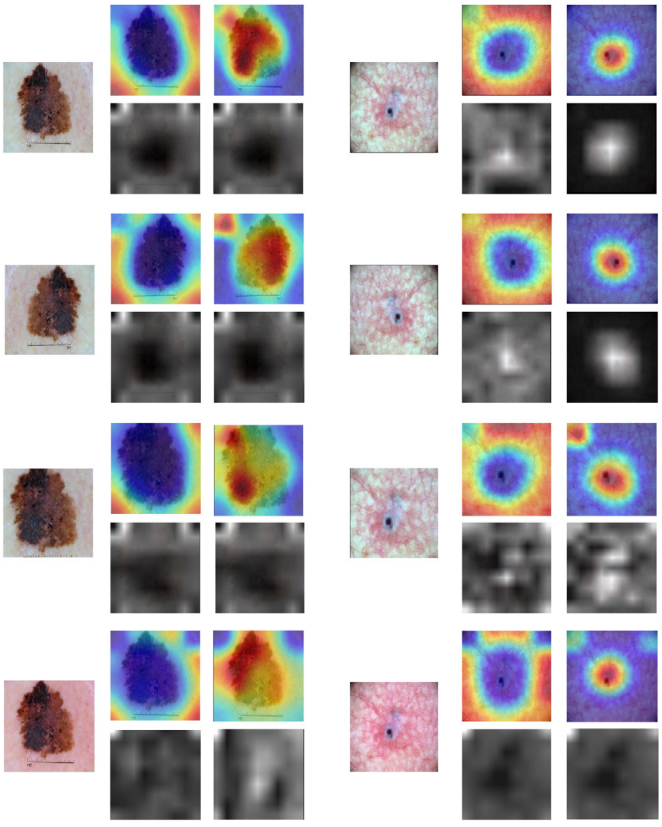

**Fig. 7.** Examples from the ISIC 2018 data set (top row) and corresponding geometric transformations: horizontal flip (2nd row), 80% central crop (3rd row), and without color normalization (4th row). The first lesion is consistently diagnosed as "Melanocytic Melanoma", while the second is always diagnosed as "Non-Melanocytic", but as "BCC" for the first two cases and as "Vascular" in the last two transformations. We also show the spatial attention maps and the most probable activation map for each of the decisions.

## 7. Conclusions

This paper proposes a diagnostic model for dermoscopy images that: (i) uses a multi-task network to perform a hierarchical diagnosis of skin lesions; and (ii) provides visual information to explain the diagnosis. By leveraging these two factors, we achieved competitive results on two state-of-the-art dermoscopy data sets (ISIC 2017 and 2018), without the need to augment the training data with external or artificially generated data and without using CNN ensembles. The experimental results show that the model can identify clinically relevant regions in the images and use them to provide a diagnosis. Additionally, the model explainability helps understand how changes in the viewpoint influence classification performance. Both factors also reveal new directions of research that can make CNNs safer to be applied in clinical practice.

In future work, we would like to extend our model to interpret, from a medical perspective, the regions highlighted by spatial attention. Additionally, we would like to improve the robustness of the features extracted by the CNNs, to make the CADS fully invariant to acquisition conditions.

## Acknowledgments

This work was supported by the FCT project and multi-year funding: [CEECIND/ 00326/2017], [PTDC/ EEIPRO/0426/2014], LARSyS - FCT Plurianual funding 2020–2023.

The Titan Xp used in this project was donated by the NVIDIA Corporation.

## References

[1] R.L. Siegel, et al., Cancer statistics, 2019, CA Cancer J. Clin. 69 (1) (2019) 7–34.
[2] M.E. Celebi, et al., Dermoscopy image analysis: overview and future directions, IEEE JBHI 23 (2) (2019) 474–478.
[3] C. Barata, et al., A survey of feature extraction in dermoscopy image analysis of skin cancer, IEEE JBHI 23 (3) (2019) 1096–1109.

[4] P. Tschandl, et al., The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (2018) 180161.
[5] A. Esteva, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.
[6] Y. Fujisawa, et al., Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis, Br. J. Dermatol. 180 (2) (2019) 373–381.
[7] G. Litjens, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[8] B. Zhou, et al., Learning deep features for discriminative localization, in: Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[9] R.R. Selvaraju, et al., Grad-CAM: visual explanations from deep networks via gradient-based localization, in: International Conference on Computer Vision, 2017, pp. 618–626.
[10] J. Gu, et al., Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.
[11] B. Zhou, et al., Interpreting deep visual representations via network dissection, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2018) 2131–2145.
[12] I. Rafegas, et al., Understanding trained CNNs by indexing neuron selectivity, Pattern Recognit. Lett. (2019).
[13] N. Sarafianos, et al., Curriculum learning of visual attribute clusters for multi–task classification, Pattern Recognit. 80 (2018) 94–108.
[14] J. Zhang, et al., Learning multi-layer coarse-to-fine representations for large-scale image classification, Pattern Recognit. 91 (2019) 175–189.
[15] G. Argenziano, et al., Interactive Atlas of Dermoscopy, EDRA Medical Publishing & New Media, 2000.
[16] J. Yang, et al., Classification for dermoscopy images using convolutional neural networks based on region average pooling, IEEE Access 6 (2018) 65130–65138.
[17] J. Zhang, et al., Attention residual learning for skin lesion classification, IEEE TMI 38 (9) (2019) 2092–2103.
[18] P. Simoens, B. Dhoedt, Visualizing convolutional neural networks to improve decision support for skin lesion classification, in: MLCN 2018, DLF 2018, and iMIMIC 2018, 11038, Springer, 2018, pp. 115–123.
[19] M. Yu, et al., Facial expression recognition based on a multi-task global-local network, Pattern Recognit. Lett. (2020).
[20] J. Kawahara, et al., 7-point checklist and skin lesion classification using multi–task multi-modal neural nets, IEEE JBHI 23 (2) (2019) 538–546.
[21] I. González-Díaz, Dermaknet: incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis, IEEE JBHI 23 (2) (2019) 547–559.
[22] P. Tschandl, et al., Accuracy of the first step of the dermatoscopic 2-step algorithm for pigmented skin lesions, Dermatol. Pract. Concept. 2 (3) (2012) 43–49.
[23] Y. Guo, et al., CNN-RNN: a large-scale hierarchical image classification framework, Multimed. Tools Appl. 77 (8) (2018) 10251–10271.
[24] D. Roy, et al., Tree-CNN: a hierarchical deep convolutional neural network for incremental learning, Neural Netw. 121 (2020) 148–160.
[25] D. Nguyen, T. Okatani, Multi-task learning of hierarchical vision-language representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10492–10501.
[26] P. Kinghorn, et al., A hierarchical and regional deep learning architecture for image description generation, Pattern Recognit. Lett. 119 (2019) 77–85.
[27] K. Shimizu, et al., Four-class classification of skin lesions with task decomposition strategy, IEEE TBME 62 (1) (2015) 274–283.
[28] C. Barata, J.S. Marques, Deep learning for skin cancer diagnosis with hierarchical architectures, ISBI, 2019.
[29] X. Shu, et al., Hierarchical long short-term concurrent memory for human interaction recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2019).
[30] C. Barata, et al., Deep attention model for the hierarchical diagnosis of skin lesions, Computer Vision and Pattern Recognition Workshops, 2019.
[31] S. Bai, S. An, A survey on automatic image caption generation, Neurocomputing 311 (2018) 291–304.
[32] O. Vinyals, et al., Show and tell: a neural image caption generator, in: Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[33] K. Xu, et al., Show, attend and tell: beural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.
[34] L. Chen, et al., SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning, in: Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).
[36] G. Huang, et al., Densely connected convolutional networks., in: Computer Vision and Pattern Recognition, 1, 2017, p. 3.
[37] K. He, et al., Deep residual learning for image recognition, in: Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[38] S. Woo, et al., CBAM: convolutional block attention module, in: ECCV, 2018, pp. 3–19.
[39] J.K. Chorowski, et al., Attention-based models for speech recognition, in: Neurips, 2015, pp. 577–585.
[40] N. Codella, et al., Skin lesion analysis toward melanoma detection: achallenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC), in: ISBI, 2018, pp. 168–172.
[41] C. Barata, et al., Improving dermoscopy image classification using color constancy, IEEE JBHI 19 (2015) 1146–1152.
[42] F. Perez, et al., 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, Springer, 2018, pp. 303–311.
[43] Y. Cui, et al., Class-balanced loss based on effective number of samples, in: Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.
[44] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980 (2014).
[45] J. Pennington, et al., Glove: global vectors for word representation, in: EMNLP, 2014, pp. 1532–1543.
[46] B. Harangi, Skin lesion classification with ensembles of deep convolutional neural networks, Elsevier JBI 86 (2018) 25–32.
[47] A. Mahbod, et al., Fusing fine-tuned deep features for skin lesion classification, CMIG 71 (2019) 19–29.

**Catarina Barata** received B.Sc. and M.Sc. degrees in Biomedical Engineering, and Ph.D. degree in Electrical and Computer Engineering from Instituto Superior Técnico, University of Lisbon, in 2009, 2011, and 2017 respectively. Currently, she is an Invited Assistant Professor at Instituto Superior Técnico, and a Research assistant at Institute for Systems and Robotics (ISR). Her research interests include the development of interpretable classification methods with applications in medical image analysis and surveillance.

**M. Emre** Celebi received the B.Sc. degree in computer engineering from the Middle East Technical University, Ankara, Turkey, in 2002, and the M.Sc. and Ph.D. degrees in computer science and engineering from The University of Texas at Arlington, Arlington, TX, USA, in 2003 and 2006, respectively. He is currently Professor and Chair of the Department of Computer Science, University of Central Arkansas, Arkansas, USA. He has pursued research in the field of image processing and analysis.

**Jorge S. Marques** received the E.E., Ph.D., and Aggregation degrees from the Technical University of Lisbon, Lisbon, Portugal, in 1981, 1990, and 2002, respectively. He is currently a Full Professor with the Electrical and Computer Engineering Department, Instituto Superior Técnico, Lisbon, and a Researcher at the Institute for Systems and Robotics. His research interests include the areas of image processing and pattern recognition. Dr. Marques was Co-chairman of the IAPR Conference IbPRIA 2005, and President of the Portuguese Association for Pattern Recognition (20012003).