uc3m | Universidad Carlos III de Madrid

Master Degree in Information and Health Engineering at UC3M
Academic Year 2019-2020

*Master Thesis*

# Learning Sequences of Dermoscopic Attributes for Explainable Skin Lesion Diagnosis

Belén Esteve Cogollos

Tutor
Miguel Ángel Fernández Torres

Madrid, October 2020

# Learning Sequences of Dermoscopic Attributes for Explainable Skin Lesion Diagnosis

Belén Esteve-Cogollos and Miguel-Ángel Fernández-Torres

**Abstract**—Computer-Aided Diagnosis Systems are getting more and more attention due to their encouraging, fast and accurate results, which allow dermatologists to have a first insight about the severity of skin lesions. However, it is difficult for the specialists to trust in these tools, since they cannot explain how they made a concrete decision. This article presents a system for skin lesion diagnosis which is able not to only detect melanomas, but also to determine a sequence of gazes over the dermoscopic image of the lesion, with the ultimate goal of offering explainability to its results. For that purpose, the model incorporates a convolutional LSTM-based attention mechanism, which is trained to recurrently segment dermoscopic attributes, in an attempt to simulate the visual attention of an expert when performing this complex task. Therefore, the system can be understood as a convolutional encoder-decoder architecture, which is composed of two branches, one for the recurrent segmentation of dermoscopic attributes and other for skin lesion recognition, based on the sequence of fixated attributes. A multi-task loss function is proposed to train this model, which integrates data balancing strategies. Different configurations and hyperparameters are tested by considering the ISIC 2017 database. Although the model possibilities have not been exhaustively explored, the configuration with the best segmentation results acquires reasonable diagnosis scores, but poor structure segmentations, being the model clearly biased to frequent structures. Therefore, for future research, using a more balanced database in terms of structure contents and/or adapting the balancing method proposed, at least for the classification of structures, are needed.

**Index Terms**—skin lesion diagnosis, dermoscopic attributes, task-driven attention mechanisms, explainable AI, convolutional encoder-decoder networks, Long Short-Term Memory units

## 1 Introduction

MElanoma recognition has lately become a topic of major concern since the number of diagnosed cases is continuously increasing [1]. Indeed, it is in the top 10 position ranking of the most common cancers for both men and women [2]. Moreover, melanoma has a low survival rate for patients in advanced stages, which enhances the importance of its early detection.

Computer-Aided Diagnosis (CAD) systems based on Deep Learning methods are getting more and more attention in this field due to their encouraging, fast and accurate results, together with the promotion of the *International Skin Imaging Collaboration* (ISIC) challenges [3]–[5]. These challenges are held every year since 2016, with the main purpose of tackling the lesion classification problem. Besides, CAD systems have become priceless diagnosis tools for dermatologists due to their objectivity, a property that cannot be applied to these experts, for whom their diagnoses could be affected by feelings and emotions depending on the day, as it is shown in [6].

Nevertheless, the major problem of most of the existing CAD systems is their lack of explainability. Indeed, dermatology specialists do not trust these tools, since the outcomes of conventional CADs directly provide the diagnosis without giving any explanation. However, some Deep Learning scientists are starting to tackle this problem by introducing new modules inside the networks. These modules are tailored to provide other useful outcomes, such as dermoscopic structures [7] and/or attention maps [8], [9]. Moreover, *ISIC challenge 2017* also introduced the lesion segmentation and feature detection tasks. Effectively, although the classification is understood as being the most important problem to tackle, its combination with one or both of the other two tasks could give rise to a more reliable and complete diagnosis tool, which not only could classify the lesion but also provide extra information, with the ultimate goal of, at least, partially explaining the classifier decision.

Following the previous statements, we propose a new architecture which performs not only a lesion diagnosis but also the extraction of a sequence of attention or gazes, aiming to improve the explainability of the network. The article makes the following contributions:

1) We propose an architecture for explainable skin lesion recognition, which is composed of a convolutional LSTM-based attention mechanism. This system has two branches, one designed to obtain dermoscopic structures information and the other to perform lesion classification. The system can be understood as a convolutional encoder-decoder with two different modules: a classifier for the diagnosis branch and a decoder for the recurrent segmentation of dermoscopic structures.
2) We make use of a multi-task loss function, which allows the system to learn either to recurrently segment dermoscopic attributes or to detect melanomas. Information related to dermoscopic at-

• *Universidad Carlos III de Madrid, Leganés, 28911, Spain.*
*E-mail: belen.esteve.co@gmail.com*
*E-mail: migferna@pa.uc3m.es*

*October 31, 2020.*

tributes allows the attention mechanism to provide at its output sequences of gazes associated to attributes, which facilitates the explainability of its recognition results.

3) We present several experiments for an in-depth analysis of the model proposed, trying different configurations and quantitatively evaluating their results both in terms of skin lesion recognition and recurrent dermoscopic attributes segmentation. Last but not least, an error analysis is performed to qualitatively assess the explainability of our approach.

The rest of the paper is organized as follows. Section 2 gathers state-of-the-art information, mainly related to skin lesion classification, dermoscopic attributes segmentation and attention mechanisms. Section 3 presents in detail the architecture proposed and the training procedure, where the multi-task loss function is explained. Section 4 describes the experiments performed to determine the best configuration for the system proposed in terms of skin lesion classification, dermoscopic attributes segmentation and explainability based on sequences of attention. Finally, Section 5 states the conclusions and future lines of research.

## 2 Related work

In this section, relevant state-of-the-art concepts and architectures related to skin lesion classification, dermoscopic attributes segmentation and attention mechanisms are briefly summarized.

### 2.1 Classification of skin lesions

Since 2016, the ISIC challenges [3] have been held to tackle the skin lesion diagnosis, leading to a wide range of different systems. The winner of the ISIC challenge 2017 [4] was the system for melanoma classification presented by Menegola *et al.* [10], which makes use of *ResNet-101* [11] and *Inception-v4* [12] as backbone networks. Despite the use of such deep networks boosted their results, they also needed, at the same time, much more computational time and memory resources than other lighter networks such as *VGG-16* [13]. Authors also tried different weighting methods to deal with data imbalance but, finally, none of them worked better than the method which did not make use of weights. However, other balancing strategies could benefit lesion classification. For the final classification, they used Support Vector Machines (SVM) [14].

Matsunaga *et al.* proposed also in 2017 a system which ended being the best average classifier. This is the one which obtained the best average AUC score over the three lesion classes (melanoma, seborrheic keratosis and nevus). Authors used a modified version of the *Resnet-50* [15], which is still deep but more computationally affordable than *ResNet-101* and *Inception-v4* models. They also used SVM for the final classification task.

Yan *et al.* used a *VGG-16* backbone network in [8], being its last classification decision made over a Fully Connected (FC) layer followed by a softmax activation function. Nevertheless, the input to the FC layer in this model is the concatenation of the activation maps obtained at different layers, some of them weighted by two attention maps provided by convolutional attention mechanisms, and after applying Global Average Pooling (GAP). The computational cost of the system is slightly higher, but the results are better. Moreover, these attention maps can be displayed in order to explain the predicted lesion class.

Although deeper networks have been proven to obtain better results, computational time is also a valuable resource. Indeed, these models are not worth the slightly better performance they acquire, since they could take twice the training time of other models. However, obtaining additional information to justify the automatic diagnosis, without drastically increasing the number of parameters of the model, could benefit this task.

### 2.2 Detection of dermoscopic structures

The dermoscopic features detection is the task that refers to locate and classify dermoscopic structures within an image of a skin lesion. This provides the expert relevant information to evaluate for the diagnosis, since the dermoscopic features depend on the type of lesion. Indeed, this information is also relevant for the lesion classification CADs and could help the network to focalize in this interesting structures. These localization of features is usually performed introducing an image in the architecture and obtaining a pixel by pixel classification. However, other methods are also used.

In [16], the encoder blocks of the *VGG-16* are used to obtain pixel-wise features, which allow the network to assign more than one class per pixel. Then, as the Ground Truths (GTs) are given in superpixels, authors also gave their predictions at superpixel level.

Lesion Feature Network (LFN) is also proposed in [17] for tackling this problem. This network has been designed from scratch, i.e., it does not consider any existing architecture. Rather than introducing the whole image into the network, inputs are superpixels, while the outputs are their corresponding labels.

Barata *et al.* introduced in [18] a new CAD system which uses the dermoscopic features to produce saliency maps. Then, these maps are introduced in an image captioning network in order to predict if the lesion is melanocytic or not, together with its particular type (melanoma, congenital nevi, atypical nevi, etc.). By obtaining these maps, which are displayable and linked with the performed diagnosis, they are introducing interpretation.

Combining dermoscopic feature extraction with displayable intermediate stages of the network could boost not only the lesion classification performance, but also the explainability of the network.

### 2.3 Sequences prediction

In Deep Learning, Recurrent Neural Networks (RNNs) based modules are used to implement many applications such as voice recognition [19] or the prediction of temporal sequences [20]. These RNNs leverage temporal information, enabling the module to share its weights for different timesteps. However, due to their 'memory' or the value of their hidden states, they tend to fail when the length of the sequence starts being too long [21], due to exploding and vanishing gradient issues. To tackle this problem, Hochreiter *et al.* [22] proposed Long Short-Term Memory (LSTM) units, which include additional gates to control the error flows caused by backpropagating very small values. Through these decades, a wide variety of LSTM cells have been proposed. In particular, the Convolutional LSTM (ConvLSTM) units [23], which take advantage not only of temporal information, but also of the spatial one. Indeed, they are made to deal with images, 2D or 3D inputs, rather than vectors.

Using this particular cell, Salvador *et al.* introduced in [24] a system for *Semantic Instance Segmentation*. Its architecture consists of a recurrent convolutional *encoder-decoder* which sequentially obtains binary masks for the objects presented in an image, together with their probabilities of belonging to a specific class. Their *encoder* is *VGG-16*-based, while the *decoder* consists of several ConvLSTMs, one at every decoder level, with the aim of dynamically capturing visual characteristics of the images at different dimensionalities. It should be noted that the generation of the sequence of masks for segmentation is due to the performance of ConvLSTM cells.

### 2.4 Attention mechanisms

Since attention mechanisms were applied to deep neural networks, they have demonstrated that they help the network to focus on the most relevant areas for the task they are tailored to manage [25]. These mechanisms were firstly used in Neural Machine Translation (NMT) [26], [27], achieving state-of-the-art results in their publication's date. Since that moment, some saliency methods have been effectively implemented. As a consequence, results have been boosted in tasks such as image classification [28]–[30] and segmentation [31], [32].

Firstly, it is important to know that there exist two types of visual attention: *pos hoc* and learnable attention [33]. The first type consists in interpreting the network parameters after it is completely trained, by making use of post-processing algorithms. The second type, in which the system introduced in this article can be located, is about introducing mechanisms inside the network, such as activation maps or gradient-based saliency methods.

Concerning learnable attention in image classification tasks, Jetley *et al.* proposed in [34] an end-to-end trainable attention module which, given the 2D feature maps of the input image, outputs a matrix of scores for each map. Later, Yan *et al.* extended this attention module in [8] to more complex non-linear operations, obtaining the attention maps by computing the summation of intermediate features and the upsampled version of global ones. Then, these maps are used as additional information in order to carry out the classification task, outperforming the results obtained using the same backbone network, but without this mechanism.

Another quite interesting learnable attention module is the one presented by Cornia *et al.* in their Saliency Attentive Model (SAM) [9]. This module, Attentive Convolutional Long Short-Term Memory (*Attentive ConvLSTM*), consists in a recurrent mechanism based on a ConvLSTM, with the particularity that it is preceded by an Attentive Model. This design enhances relevant areas of the modules' input to focus its efforts. The module is used in this paper for the refinement of saliency maps for static images, rather than to obtain a temporal sequence of gazes, as in our approach.

## 3 Explainable CAD system for skin lesion recognition

### 3.1 Architecture overview

Given the challenging task of skin lesion recognition, our approach aims to tackle at the same time the image classification, segmentation, and features detection tasks. For this purpose, we propose a convolutional encoder-decoder with two branches and a central recurrent attention mechanism . Each the branches has a concrete objective: the decoder or dermoscopic attributes segmentation branch serves for recurrent semantic segmentation, while the diagnosis branch solves skin lesion recognition by making use of the sequences of visual attention maps generated for the purpose of the first decoder.

Therefore, as is represented in the diagram of Figure 1, the proposed architecture firstly consists of the three initial *VGG-16* [13] CNN blocks as the encoder. Then, an *Attentive ConvLSTM* [9] module is attached, which constitutes the core of the system and provides sequences of gazes for the explainability of its results. Lastly, the two branches for the segmentation of dermoscopic structures and the classification of the skin lesion are consecutively placed. In addition, skip connections [35] are added to introduce the information from the encoder directly to the decoder layers, which helps eliminating the singularities generated by the non-identifiability of the model.

On the one hand, the branch for the recurrent segmentation of the lesion and the dermoscopic attributes consists of a convolutional decoder, which produces skin and dermoscopic features binary masks at different timesteps, also classifying them according to their classes. On the other hand, the diagnosis branch basically consists in the continuation blocks of the *VGG-16* [13] followed by a classifier, which outcomes the type of lesion. The different parts of the network are explained in detail in the following subsections.

#### *3.1.1 Backbone architecture: Encoder and attention mechanism*

The backbone or common part of the network to deal with the segmentation of structures and classification tasks is composed of an encoder and a *Attentive ConvLSTM* [9] module.

The encoder is composed of the first three blocks of the batch-normalized version of the *VGG-16* network as in [8], which has been the starting point of this work. It receives a RGB pre-processed image of the lesion $X \in \mathbb{R}^{224 \times 224 \times 3}$. At the end of these blocks, the output is a set of convolutional activation maps $X' \in \mathbb{R}^{28 \times 28 \times 256}$. $X'$ is introduced into the *Attentive ConvLSTM* module, which is identical to the one used in SAM [9] and described here for the sake of completeness. This module is the key piece of our network, which enables the production of different outcomes in the same execution for a range of time instants, giving rise to a sequence of gazes for the explainability of the skin lesion classification task. This sequence of gazes is based on a sequence of $T$ spatial attention maps $A_t$ generated by this module, *i.e.* the sequence of attention $A_t$, $t \in [0, T-1]$, which highlight the regions where the network is focusing on at each timestep, in order to finally provide a diagnosis, determining if a skin lesion is a melanoma or not.

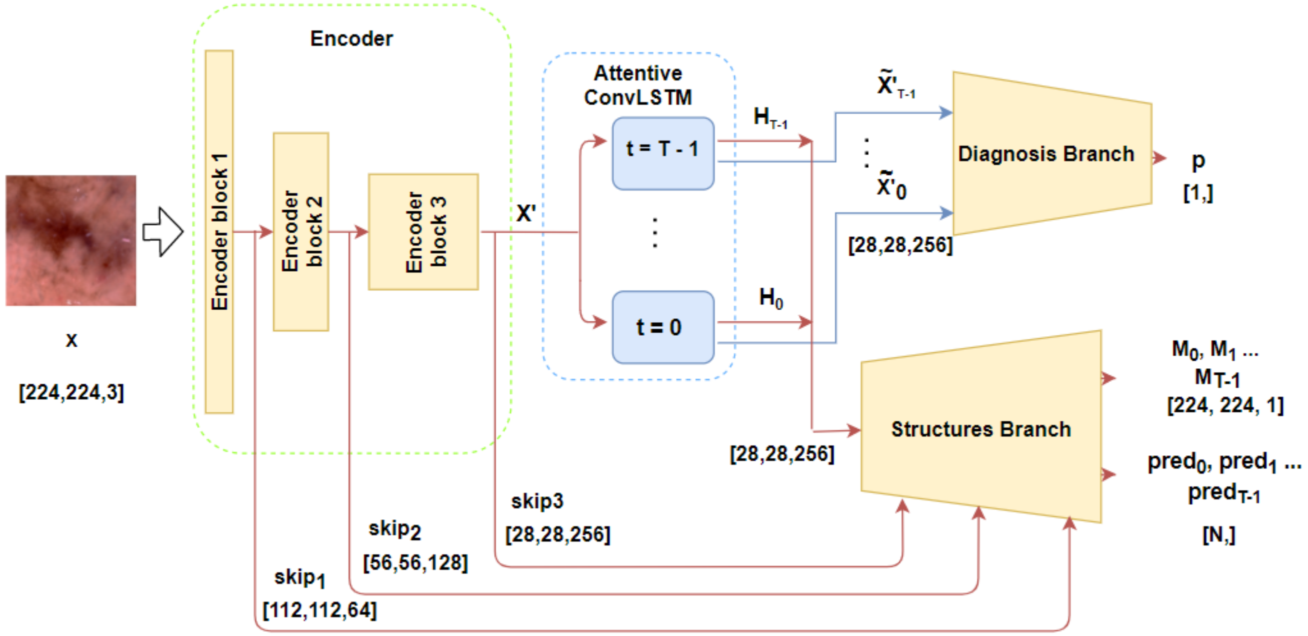

Fig. 1. Simplified scheme of the proposed architecture for explainable skin lesion recognition. All the modules are displayed, the input $X$ and the outputs: probability associated to diagnosis $p$, obtained binary segmentation masks for dermoscopic structures at each timestep $M_{0,1...T-1}$ and their corresponding output probabilities $pred_{0,1...T-1}$

.

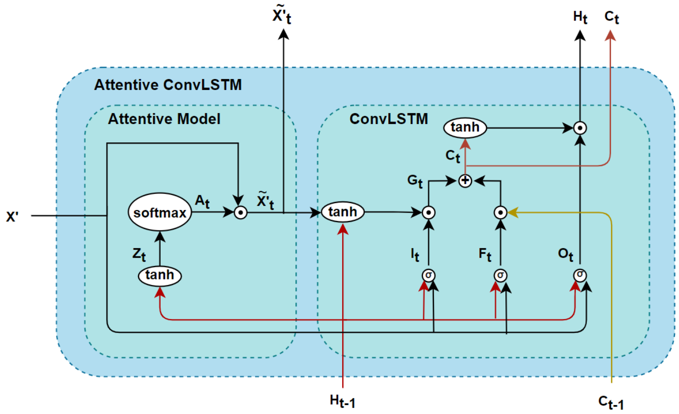

Fig. 2. The Attentive ConvLSTM cell [9] used for the attention mechanism of the system proposed. X′ is the activation map obtained from the encoder module, $H_t$ and $H_{t-1}$ the hidden states and $C_t$ and $C_{t-1}$ the memory cells.

The attention mechanism defined by the *Attentive ConvLSTM* module (see Figure 2) is divided into two main parts, the *Attentive Model* and the *Convolutional LSTM*. The first one obtains an attention map $A_t \in \mathbb{R}^{28\times 28}$ at each timestep. In order to obtain $A_t$, $X'$ and the ConvLSTM hidden state $H_t$ are convolved by $W_a$ and $U_a$, respectively. Then, the result of these convolutions is summed and introduced into a hyperbolic tangent ($tanh$) function. Lastly, it is convolved with a one-dimensional convolutional kernel $V_a$ as follows:

$$Z_t = V_a * tanh(W_a * X' + U_a * H_{t-1} + b_a) \qquad (1)$$

After, a $softmax$ function is applied to obtain the attention map $A_t$. Therefore, for each spatial location or image coordinates $(i, j)$, the value of attention is computed as follows:

$$A_t^{ij} = p(attention_{ij}|X', H_{t-1}) = \frac{exp(Z_t^{ij})}{\sum_i \sum_j exp(Z_t^{ij})} \qquad (2)$$

Finally, $A_t$ is element-wise multiplied to each channel in $X'$, which produces activation maps $\widetilde{X'_t}$ weighted by attention, highlighting the most conspicuous local features of the lesion:

$$\widetilde{X'} = A_t \odot X' \qquad (3)$$

Moreover, the obtained attention maps $A_t, t \in [0, T-1]$ provide valuable and visible spatial information, which could partially clarify the posterior decisions of the network, as it is discussed later in Section 4.

The second stage of the attention mechanism is the *ConvLSTM* layer. It sequentially updates its hidden state $H_t$ by using three gates ($I_t$, $F_t$ and $O_t$), a candidate memory $G_t$ and the current and previous memory cells ($C_t$ and $C_{t-1}$), as it is defined in Eqs. 4-9.

$$I_t = \sigma(W_i * \widetilde{X'_t} + U_i * H_{t-1} + b_i) \qquad (4)$$

$$F_t = \sigma(W_f * \widetilde{X'_t} + U_f * H_{t-1} + b_f) \qquad (5)$$

$$O_t = \sigma(W_o * \widetilde{X'_t} + U_o * H_{t-1} + b_o) \qquad (6)$$

$$G_t = tanh(W_g * \widetilde{X'_t} + U_g * H_{t-1} + b_g) \qquad (7)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot G_t \qquad (8)$$

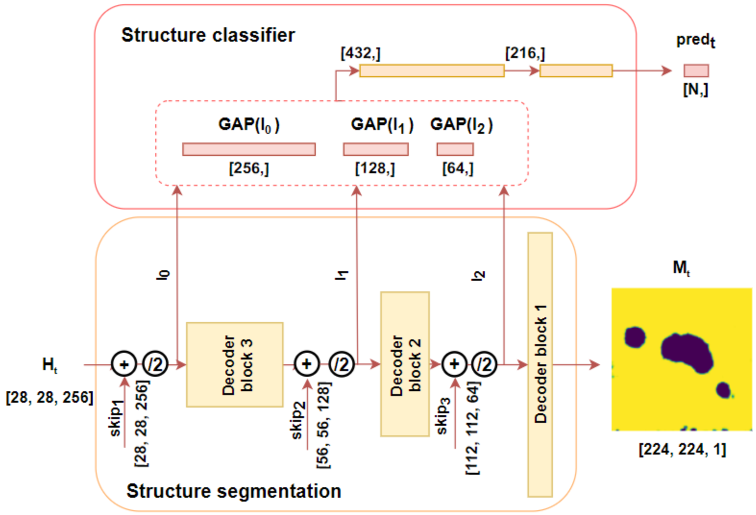

Fig. 3. Diagram of the branch for recurrent segmentation of dermoscopic structures. The inputs of this part are the hidden state $H_t$ and the skip connections of the corresponding encoding levels. For the structure classifier, the concatenated global poolings of $l_0$, $l_1$ and $l_2$ are introduced into a sequence of two FC layers.

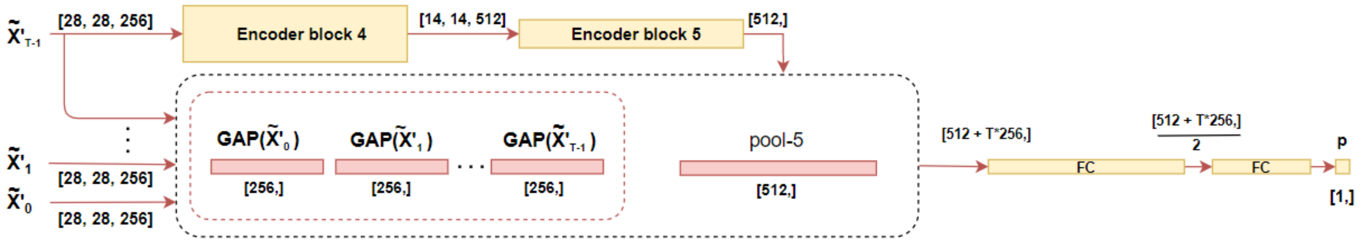

Fig. 4. Diagram of the branch for skin lesion classification. The inputs of this branch are the activation maps weighted by attention $\widetilde{X'}_t$. The $\widetilde{X'}_{T-1}$ is passed through the last decoder blocks and then concatenated with the GAP versions of the rest of the $\widetilde{X'}_t$.

$$H_t = O_t \odot tanh(C_t) \tag{9}$$

At the end of each iteration, the hidden state $H_t \in \mathbb{R}^{28\times28\times256}$ is delivered to the decoder, enabling it to produce structure binary masks and determine their corresponding classes at each timestep. After, at the output of the Attentive ConvLSTM, all the generated $\widetilde{X'}_t$, are passed through a GAP operation and then concatenated. This vector, together with the last activation maps obtained ($\widetilde{X'}_{T-1}$), also processed by a GAP operation, are now used as input to the classifier at the top layer of the model proposed.

### 3.1.2 Branch for recurrent segmentation of dermoscopic structures

The decoder branch is represented in Figure 3 and has been built to segment and classify different spatial regions of the skin lesion images according to several dermoscopic attributes. Indeed, the pursued benefit of this decoder is twofold: 1) the obtention of interpretable information for the dermatologist and, indirectly, 2) the boost in the performance of the diagnosis branch due to the joint training of the two branches of the network.

The module for segmentation is a *VGG-16 decoder*, similar to the one presented in *SegNet* [36], which is symmetric to the *encoder*. It consists of three 2D convolutional blocks, batch normalization layers and ReLUs for non-linearities. Unlike the *SegNet decoder*, which uses unpooling for the dimensionality augmentation, each block is preceded by an upsampling of scale factor equal to 2. As a consequence of this operation, receiving an input of shape $[28, 28, 256]$ results in binary masks of dimension $[224, 224, 1]$ for the recurrent segmentation of dermoscopic attributes task. In addition, the skip connections are implemented by averaging the encoder and decoder activation maps right before each decoder block. As a result, the input of the next decoder block is a set of activation maps which covers the encoder and decoder informations.

Given the binary mask classification task, the GAPs are computed before each decoder block. After they are concatenated. As a result, a vector of length $448$ $(256 + 128 + 64)$ is obtained. Then, this vector passes through two dense layers, obtaining a vector of the same length as the total number of possible dermoscopic structures. This last vector feeds a *softmax* function, resulting in a vector of probabilities of belonging to each of the possible structure classes considered.

### 3.1.3 Branch for skin lesion classification

The aim of the branch for skin lesion classification is to predict if dermoscopic images contain melanomas or not. Hence, as it is shown in Figure 4, the dimensionality of the maps has to be drastically reduced from the activation maps dimensions $[28, 28, 256]$ to the number of classes, which is two (*benign* and *melanoma*). Thus, the classifier consists of the last encoder *VGG-16* blocks, particularly blocks *pool4* and *pool5*, achieving a dimensionality of $[14, 14, 512]$. Then, after the last pooling, the activation maps have been reduced into a vector of length $512$ by applying GAP. Then, this vector is concatenated with the GAP versions of the attention-weighted activation maps $\widehat{X'}_t, t \in [0, T-1]$ obtained during the *Attentive ConvLSTM* execution. In this way, the sequence of $T$ gazes over the image, which are established for the recurrent segmentation of dermoscopic attributes, is taken into account for lesion segmentation.

## 3.2 Training

In order to efficiently train the complete architecture proposed, the following aspects must be taken into account:

1) One of our main objectives is to obtain a sequence of gazes similarly to how a dermatologist could attend to a skin lesion when offering a diagnosis. Hence, the output binary masks should not follow a strict order.
2) The network has two output branches, which means that the global loss should have at least two terms (diagnosis and structures). Furthermore, the decoder branch for the recurrent segmentation of dermoscopic attributes tackle both the structure segmentation and classification tasks. Therefore, the structure loss also consists of two terms.
3) Not all the lesions contain the four dermoscopic structures considered in our experiments, according to ISIC 2017 [4], which are pigment network, negative network, milia-like cysts and streaks. In fact, the majority of them present just a single dermoscopic structure. Hereafter, the lesion and skin masks are going to be treated as two additional structures that could be selected by the gazes of the network. The database is further analyzed in Section 4.1.

The different strategies to deal with these issues are explained in the next subsections.

*Unsorted sequences of dermoscopic attributes*

Since our structure outcomes are not ordered, the GT assignation to each of the output masks is not a trivial task. For this purpose, the soft Intersection over the Union (softIoU) (see Eq. 10) is used as a measure of similarity between predicted $\hat{M}$ and GT $M$ masks, as in [24].

$$softIoU(\hat{M}, M) = 1 - \frac{\langle \hat{M}, M \rangle}{\|\hat{M}\|_1 + \|M\|_1 + \langle \hat{M}, M \rangle} \quad (10)$$

First, a matrix of comparisons is build by computing the softIoU for each pair of predicted-GT masks. As a result, a matrix of size $S \times T$ is obtained, where $S$ is the total number of structures considered and $T$ the number of timesteps in the attention sequence. Then, the Hungarian algorithm [37] is used to select the most similar predicted-GT pairs of masks.

The Hungarian algorithm is an optimization procedure which, when being applied to a matrix, outcomes the permutations in rows and columns in order to obtain its minimum values at the principal diagonal. Note that each row and column can only appear once in the matrix. Hence, each GT mask will be assigned to a different predicted mask and vice versa. By using this algorithm, we are also forcing the network to learn other masks different from the wider and most common structures considered (*lesion* and *skin*).

In contrast, GT classes for structures are obtained in the evaluation stage by pairing each output mask with the GT one for which the softIoU score obtained is higher. Thus, at this stage, GT masks could be used repeatedly or even not appear at all.

*Multi-task loss function*

Here we describe the multi-task loss function considered for training the model proposed. First, given the branch for skin lesion classification, the selected criterion $\mathcal{L}_{diag}$ is the Focal Loss (FL) [38]. This decision was made based on the experimental results obtained by Yan *et al.* in [8] and the fact that it down-weights the most frequent class samples on the loss. As it is shown in Eqs. 11 and 12, where $y_n$ is the GT label for the lesion $n$ ($y_n = 0$ for benign lesion and $y_n = 1$ for melanoma) and $\hat{y}_n$ the probability of being a melanoma, the only difference between the Binary Cross Entropy (BCE) and the FL is the multiplication by the $(1 - \hat{y}_n)^\theta$ and $\hat{y}_n^\theta$ terms. When $\theta = 0$, the FL and the BCE are the same. However, for $\theta > 0$, the loss is focused more on the missclassified samples by decreasing the relative loss of the well-classified samples.

$$BCE(y_n, \hat{y}_n) = -y_n log(\hat{y}_n) - (1 - y_n) log(1 - \hat{y}_n) \quad (11)$$

$$FL(y_n, \hat{y}_n) = -y_n(1 - \hat{y}_n)^\theta log(\hat{y}_n) - (1 - y_n)\hat{y}_n^\theta log(1 - \hat{y}_n) \quad (12)$$

Second, regarding the branch for recurrent segmentation of dermoscopic structures, we make use of two additional loss terms. On the one hand, the 1-softIoU [39] score is used for the computation of the segmentation loss term $\mathcal{L}_{seg}$. This term is obtained by calculating the $softIoU$ scores (see Eq. 10) over the GT-predicted pairs of masks, obtaining one score for each structure considered. Then, the total $\mathcal{L}_{seg}$ is the average of these obtained scores.

On the other hand, the loss considered for the structures classification task $\mathcal{L}_{class}$ is the *Categorical Cross Entropy* (CCE), which is defined for each timestep as follows:

$$CCE(y_s, \hat{y}_s) = -\sum_{s=1}^{S} y_s \cdot log(\hat{y}_s) \quad (13)$$

where $S$ is the total number of considered structures, $\hat{y}_s$ is the probability of belonging to the class $s$ and $y_s = 1$ if $s$ corresponds to the true structure class. Similarly to $\mathcal{L}_{seg}$, the final term is obtained by averaging the $CCEs$ computed over the GT-predicted pairs of classes.

Taking all these loss terms into account, the multi-task loss function defined is a weighted sum of the different terms and can be expressed as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{diag} + \beta(\gamma\mathcal{L}_{seg} + \mu\mathcal{L}_{class}), \quad (14)$$

being $\alpha$, $\beta$, $\gamma$ and $\mu$ the different weights associated to each loss term, which are empirically determined and subjected to the following conditions: $\alpha + \beta = 1$ and $\gamma + \mu = 1$.

*Unbalanced dermoscopic structures data*
The ISIC 2017 database [4] is imbalanced in dermoscopic structures. Firstly due to their nature, because the structures do not tend to occupy the same percentage of pixels in the image. Lastly, because the number of images which contain each of the structures varies tremendously. Data imbalance is further analyzed in Section 4.1.

In order to deal with the imbalance of dermoscopic structures, we propose three weighting techniques. These strategies have in common that they multiply each of the structures losses by different weights.

1) Weighting based on the presence of each structure in the training set of the ISIC 2017 [4]. The weights are computed by considering the number of images of the training set $\mathcal{P}_s$ for which each structure appears. The presence of each structure in the training set is shown in Table 1. Given each structure $i$, its associated weight $w_i$ is computed as the sum of the presences of the remaining structures $s \neq i$, which is normalized by dividing between the sum of all the presences. Eq. 15 determines this computation:

$$w_i = \frac{\sum_{s \neq i} \mathcal{P}_s}{\sum_{s=1}^{S} \mathcal{P}_s}, \tag{15}$$

where $S$ is the total number of dermoscopic structures considered. Following this equation, the most frequent structures in the database are receiving a smaller weight.

2) Another weighting strategy based on the presences of the structures in the database. Given each structure $i$, its associated weight $w'_i$ is computed as the inverse of the structure presence $\mathcal{P}_i$ multiplied by the sum of the inverse presences of all the structures $s \in [1, S]$. Eq. 16 determines this computation:

$$w'_i = \frac{1}{\mathcal{P}_i \sum_{s=1}^{S} \frac{1}{\mathcal{P}_s}} \tag{16}$$

where $S$ is the total number of dermoscopic structures considered. Following this equation, the most frequent structures in the database are receiving a smaller weight.

3) Mini-batch balancing, similarly to the method used by Kawahara *et. al* in [16]. It consists of averaging the structure loss terms based on the times they appear in a batch. Thus, frequent and infrequent structures will have the same weighting, no matter how many times they appear in the batch.

As methods 1 and 2 may look similar, their corresponding weights are shown in Table 2, evidencing that, in practice, are quite different.

## 4 Experiments and Results

In this section, the results obtained for the different tasks and configurations of our approach are shown and analyzed. The following subsections cover the experimental design, the quantitative results for the skin lesion classification and dermoscopic feature detection tasks, and a qualitative error analysis regarding the explainability extracted from the ConvLSTM-based attention mechanism and the sequences of gazes provided by the model proposed.

| Structure | *Presence in $n^o$ of images* ($\mathcal{P}_{struct}$) | *Presence in images* (%) | *Presence in pixels* (%) |
|---|---|---|---|
| *Lesion* | 2000 | 100.00 | 14.83 |
| *Skin* | 2000 | 100.00 | 80.95 |
| *Pigment network* | 1131 | 56.55 | 3.64 |
| *Negative network* | 126 | 6.30 | 0.19 |
| *Milia-like cysts* | 573 | 28.65 | 0.35 |
| *Streaks* | 116 | 5.80 | 0.07 |

TABLE 1: *Presence of the structures within the training set of the ISIC 2017 database [4]. The total sum of the values of the columns is greater than 100 because the images could contain more than one structure, as well as pixels sometimes are labeled according to more than one of the four dermoscopic features considered. Third and fourth columns are rounded to the second decimal*

| | *Lesion* | *Skin* | *Pigment network* | *Negative network* | *Milia-like cysts* | *Streaks* |
|---|---|---|---|---|---|---|
| *(2)* $w_i$ | 0.133 | 0.133 | 0.162 | 0.196 | 0.181 | 0.196 |
| *(1)* $w'_i$ | 0.025 | 0.025 | 0.044 | 0.394 | 0.086 | 0.427 |

TABLE 2: *Different weights for the two weighting strategies based on the presence of the structures in the training set of the ISIC 2017 database [4]. Method (1) penalizes harder the most frequent structures than method (2). Their values are rounded to the third decimal.*

### 4.1 Experimental design

*Database*

Images used for the experiments presented in this article constitute the ISIC 2017 Challenge [4] database. The database is composed of 2000 training, 150 validation and 600 test images, together with their corresponding GT class labels, lesion masks, superpixels images, *json* files which contain the annotation of dermoscopic features at pixel-level and metadata (age and sex). Nevertheless, we have made use of the visual data, discarding the metadata information. The ISIC 2017 database was selected for two reasons: first, because of the spatially annotated dermoscopic features and, second, due to the high amount of networks in the literature trained using it as benchmark, which enables the comparison of results.

On the one hand, this database classifies the images within three classes: *melanoma*, *seborrheic keratosis*, and *nevus*, being these last two benign types of lesion. Hence, taking into account the classification according to melanoma/benign-lesion, the database is highly unbalanced. Indeed, only around 19% of the training images are labeled as melanomas.

On the other hand, the dermoscopic structures annotated are *pigment network*, *negative network*, *milia-like cysts* and *streaks*. These are four of the evaluated features in the *7-point checklist* of Argenziano *et al.* [40], which is a well-known dermatologist method of diagnosis. As well as in lesion classification, the GT structures' content at pixel-level is highly unbalanced. As was introduced above, Table 1 indicates the presence of each of the structures in the database. It should be noted that, in our experiments, skin and lesion are considered as two additional structures and, as a consequence, the total number of structures is six. Having the four GT dermoscopic structure masks $(M_{pig}, M_{neg}, M_{mil}, M_{str})$ and the original GT lesion segmentation $M_{lesion}$, the skin mask $M_{skin}$ is computed through logical operations, stressing the pixels which are not labelled as lesion or structure. Eqs. 17 and 18 show how $M_{skin}$ is acquired through logical gates $OR\ (+)$ and $NOT\ (\bar{\cdot})$. Then, Eq. 19 shows the modification performed over $M_{lesion}$ to take into account the incorporated skin mask.

$$M_{dermo} = M_{pig} + M_{neg} + M_{mil} + M_{str} \tag{17}$$

$$M_{skin} = \overline{M_{lesion} + M_{dermo}} \tag{18}$$

$$M'_{lesion} = \overline{M_{skin} + M_{dermo}} \tag{19}$$

### Data pre-processing

Before introducing the images into the network, a pre-processing stage is considered, which adapts the input data to the required size and performs data augmentation. We can distinguish between two groups of pre-processing operations: a general pre-processing group, which is applied to the whole database (train, validation and test), and an additional pre-processing performed over the training set. It should be noted that all operations except for normalization are applied to both $RGB$ images and binary masks.

On the one hand, the general pre-processing group consists of the following three main steps:

- **Resizing:** the data is resized to width and height $[256, 256]$, a slightly greater size than the required one.
- **Center Cropping:** the borders are removed, resulting in the required network's input size, which is $[224, 224]$.
- **Normalization:** the values of the 3 channels in $RGB$ images are standardized to have zero-mean and one-standard deviation. For that purpose, the mean and standard deviation of the distribution of the original $RGB$ channels are computed over the training set.

On the other hand, the additional training pre-processing consists of the introduction of random rotations and flippings, in order to improve the model invariance against these transformations.

### Experimental setup

The main objective of the experiments carried out for the analysis presented in this paper is to determine the best configuration and parameters for the model proposed. With this aim, we assessed the performance of four different configurations:

- The architecture proposed, which includes skip connections and batch normalization.
- The architecture proposed, which includes skip connections and frozen batch normalization layers, which are pre-trained on ImageNet database [41].
- The architecture proposed, without skip connections and batch normalization.
- The architecture proposed, without skip connections but including batch normalization layers.

We also evaluated the three different balancing techniques over the structures loss proposed:

- Weighting each structure directly based on the presences of the structures in the database $(w'_i)$.
- Weighting each structure term based on the presences of the other structures through the database $(w_i)$.
- The mini-batch balancing.

### Evaluation metrics

When it comes to the assessment of the performance of our approach, different evaluation metrics are considered for each of the tasks it attempts to solve. First, regarding the recurrent segmentation of dermoscopic attributes task, the Mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds [42] is used. For each of the selected thresholds, this scoring method works in the following way:

1) Compute the IoU for all possible pairs of predicted and GT structure masks. As a result, a matrix of size $S \times T$, being $S$ the number of GT structure classes and $T$ the number of recurrent timesteps, is obtained.
2) Filter the obtained IoUs by taking into account the value of the threshold. Thus, each of the cells of the matrix now indicates if each of the pairs defined constitutes a miss or a match.
3) Calculate the precision for the matrix as follows:

$$precision = \frac{TP}{TP + FP + FN} \tag{20}$$

where $TP$ are the hits or correctly predicted pairs, $FP$ are predicted masks which do not match with any $GT$ mask and $FN$ are the GT masks that do not match with any of the predicted ones.

After this process has been performed for each of the pre-defined thresholds, the mAP at different IoU thresholds is the average of all the calculated precisions.

Continuing with the dermoscopic features assesment, a confusion matrix for structure classification has been obtained. This matrix visually relates the GT and predicted structures classes by displaying, for each GT structure class, the percentage of masks correctly or wrongly classified.

| Model | *AUC* | *AP* |
|---|---|---|
| *VGG* | 0.8212 | 0.5892 |
| *VGG_GAP* | 0.8661 | 0.6579 |
| *VGG_AttentiveConvLSTM* | 0.8341 | 0.5460 |
| *VGG_AttentiveConvLSTM_GAP* | 0.8477 | 0.6216 |
| *VGG_AttentiveConvLSTM_Deco_GAP_3* | 0.8291 | 0.5648 |
| *VGG_AttentiveConvLSTM_Deco_GAP_5* | 0.8256 | 0.5576 |
| *VGG_AttentiveConvLSTM_Deco_GAP_17* | 0.8473 | 0.5947 |

TABLE 3: Ablation study. Results on skin lesion classification obtained over the test set of ISIC 2017 database [4] for different variants of the model proposed.

Second, when computing the skin lesion classification score, the Area Under the ROC Curve (AUC) [43] and the Average Precision (AP) [44] are used. The AUC represents the area under a curve which relates the True Positive Rate (TPR) with respect to the False Positive Rate (FPR) (see Eq. 21) for all the possible thresholds in the range $[0, 1]$ [45]. Meanwhile, the AP is the average of the precisions [eq. 20] computed for all the possible thresholds in the range $[0, 1]$.

$$TPR = \frac{TP}{TP + FN},\ FPR = \frac{FP}{TP + FN} \tag{21}$$

### Implementation details

The system has been implemented in Python using the PyTorch library for Deep Learning. Other libraries have been used such as OpenCV for image processing, Matplotlib for visualization, Numpy for array treatment and Scikit-Learn for computing evaluation metrics. All the experiments have been carried out within the Google Colab GPU environment, for which the GPU characteristics are not granted and the assigned GPU could be different between sessions. The following options and hyperparameters have been chosen during the training stage: Adam optimizer with $(\beta_1, \beta_2) = (0.92, 0.96)$; batch size equal to 8; weigths for the multi-task loss $\alpha = 0.25$, $\beta = 0.75$, $\gamma = 0.75$ and $\mu = 0.25$; three different learning rates $lr_{encoder} = 0$ (transfer learning), $lr_{AttentiveConvLSTM} = 10^{-2}$ and $lr_{branches} = 10^{-4}$ for dermoscopic features segmentation and skin lesion classification branches; weight decay is $10^{-6}$ and the total number of epochs is 20, but saving checkpoints only when validation loss is improved. Code is publicly available on GitHub[1].

## 4.2 Results on skin lesion classification

In order to place into context the results obtained for the skin lesion classification task, an ablation study and a comparison with the state-of-the-art are carried out in this section. First, we are going to determine and analyze the effect of the different modules over the final diagnosis through the ablation study. Then, a comparison with other models in the literature is discussed.

| Model | *AUC* |
|---|---|
| *ISIC 2017's best [46]* | 0.911 |
| *DermaKNet [7]* | 0.896 |
| *VGG_AttentiveConvLSTM_Deco_GAP_3* | 0.829 |
| *VGG_AttentiveConvLSTM_Deco_GAP_5* | 0.826 |
| *VGG_AttentiveConvLSTM_Deco_GAP_17* | 0.847 |

TABLE 4: Comparison with the state-of-the-art. AUC scores for skin lesion classification obtained over the test set of ISIC 2017 database [4] for the best two methods in the literature and different variants of the model proposed.

### *Ablation study*

Table 3 summarizes the ablation study carried out for the skin lesion classification task. We consider the following models and setups:

- *VGG*: the original *VGG-16* model [4], whose top layer is adapted to solve a binary classification task.
- *VGG_GAP*: the *VGG-16* model, but replacing its top FC layers by the use of the GAP operation over the output of pool5 layer.
- *VGG_AttentiveConvLSTM*: the original *VGG-16* architecture, incorporating the Attentive ConvLSTM module.
- *VGG_AttentiveConvLSTM_GAP*: the *VGG-16* architecture, together with the Attentive ConvLSTM and considering the GAP strategy for classification.
- *VGG_AttentiveConvLSTM_Deco_GAP_3*: This is the first model which introduces the decoder for recurrent dermoscopic features segmentation. It is composed of the *VGG-16* encoder, the Attentive ConvLSTM and considers the GAP strategy for classification. This model uses the $X_T$ GAP version, as explained in Section 3.1.3. The balancing strategy used is the weighting through the batch occurrences. The skip layers and frozen batch normalization setup is used.
- *VGG_AttentiveConvLSTM_Deco_GAP_5*: It is composed of the *VGG-16* encoder with the Attentive ConvLSTM, the decoder for dermoscopic features segmentation and the GAP strategy for classification. This model uses the $X_T$ GAP version. The balancing strategy is the weighting with the $w'_i$ weights. The no-skip layers and batch normalization setup is used.
- *VGG_AttentiveConvLSTM_Deco_GAP_17*: It is composed of the *VGG-16* encoder with the Attentive ConvLSTM, the decoder for dermoscopic features segmentation and the GAP strategy for classification. This model uses the $X'$ GAP version. In the classifier, a FC layer is introduced to reduce the dimension of the concatenated $X_t$ GAPs before its concatenation with the outputs of pool5. The balancing strategy is based on the use of $w_i$ weights. The no-skip layers and batch normalization setup is used.

First, it can be observed that the AUC score obtained

[1] https://github.com/BelenEsteve/LearningSequences

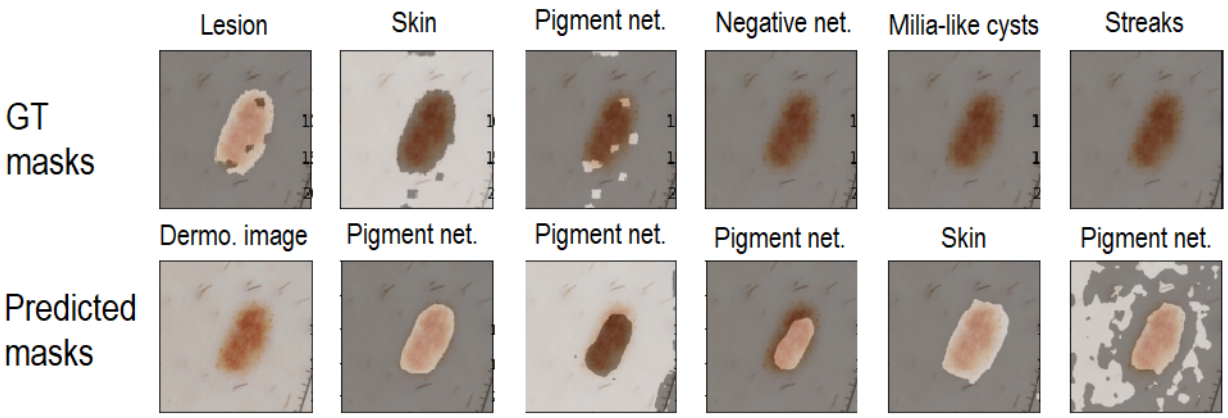

Fig. 5. Segmentation results over an example image taken from the ISIC 2017 [4] database. In this image, we can compare the GT masks (top row) with the obtained segmentations (bottom row from the second image).

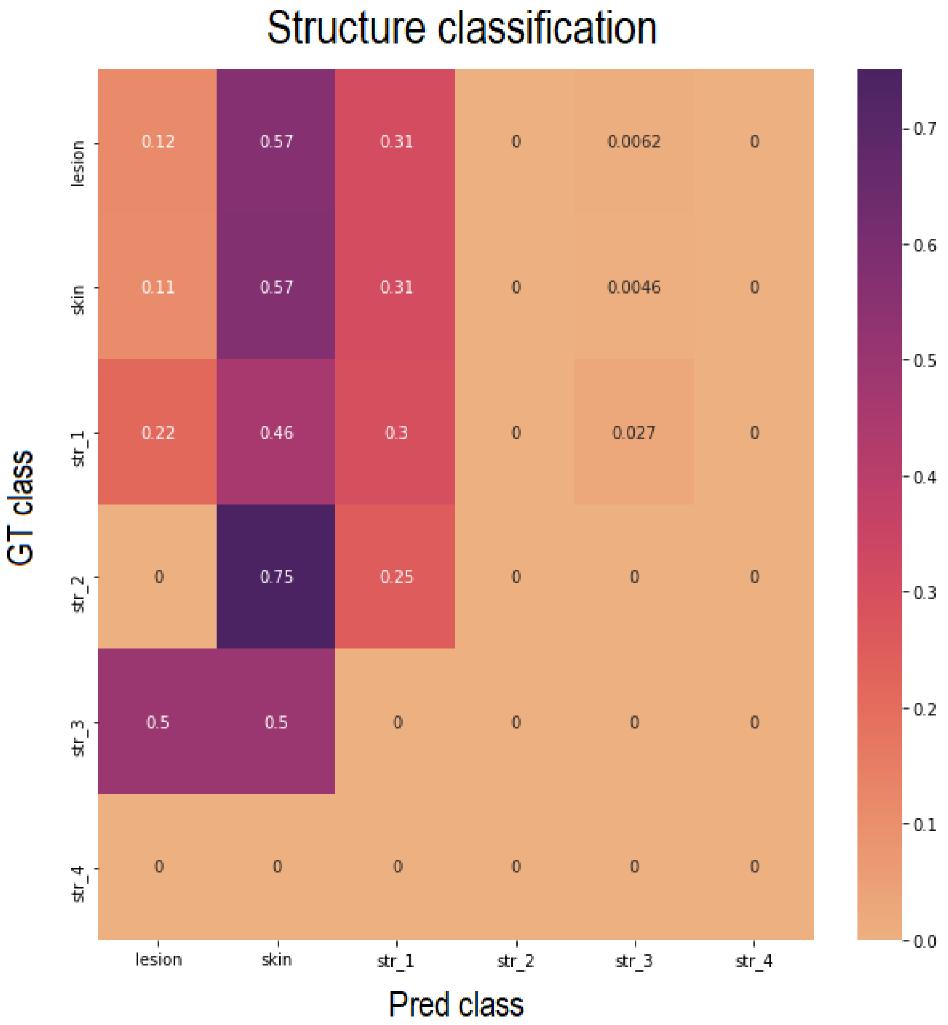

Fig. 6. Confusion matrix for the structure classification task over the test dataset of ISIC 2017 [4] database. Each cell represents the proportion of the masks of a concrete GT class which have been classified as a concrete predicted class. Hence, the row values sum up to 1. The labels mean: *str*_1 - pigment network, *str*_2 - negative network, *str*_3 - milia-like cysts, *str*_4 - streaks.

| Model | *AUC* |
|---|---|
| *Kawahara and Hamarneh[16]* | 0.895 |
| *LIN [17]* | 0.833 |
| *VGG_AttentiveConvLSTM_Deco_GAP_3* | 0.452 |

TABLE 5: Comparison with the state-of-the-art for the dermoscopic features segmentation task. Results obtained over the test set of the ISIC 2017 database [4].

by all approaches is better than the one offered by the simple *VGG* model. This could be suggesting that relevant information is being introduced by using the modules incorporated ex post. However, against our intuitions about the usefulness of the information based on dermoscopic structures, which is collected by the ConvLSTM-based attention mechanism, the *VGG_GAP* model produces the best scores in this task. This could be caused by the fact that the recurrent segmentation-based models (named as *VGG_AttentiveConvLSTM*) are generating redundant information. This might be due to the low amount of pixel-level dermoscopic features' GT annotations in ISIC 2017 database [4], which hinders learning the complex task of segmenting dermoscopic attributes. Still, the latter helps the attention mechanism to produce sequences of gazes or relevant regions in the lesion for its classification, which allow for the explainability of the diagnosis.

Another reason to justify the lower classification scores, which are the ones obtained for models which incorporate the decoder for recurrent segmentation, could be that the training procedures for the two branches of the network (diagnosis and structures) are in conflict. As the segmentation task is far more complex, its weighting at the total global loss is set to a higher value in the training stage. As a result, this could be affecting the skin lesion classification results.

*Comparison with the state-of-the-art*

When looking at Table 4, we can observe that our results are quite far from the scores of similar approaches in the literature. This might be due to these two models are designed to tackle fewer tasks. Indeed, ISIC 2017's best approach is only designed to deal with the classification task. Intuitively, as dermatologists diagnose based on dermoscopic structures, our results could be suggesting two main problems: the lack of labelled structures in the dataset and the difficulty of finding the best weights and hyperparameters for this model, due to the computational time limitations of the Google Colab GPU environment.

### 4.3 Results on recurrent segmentation of dermoscopic structures

Regarding Table 5, we could rapidly say that our models are far from the state-of-the-art scores. However, to the best of our knowledge, our models perform the first recurrent segmentation over skin lesion structures, acquiring a mAP score of 0.037. This means that further studies and experiments are needed to boost its performance.

The structure segmentation task is highly complex, even more if combined with the fact that it has to be trained jointly with the diagnosis task. Also, as the scores are computed taking as GT the mask of the predicted structure class, this is reflected in the segmentation score if the classification is wrong. This is also happening in the results of our approach, since the structure classification accuracy is 0.2203 for *VGG_AttentiveConvLSTM_Deco_GAP_3*. This may suggest that other kinds of training could have worked better, such as training the structures branch first and, after a few epochs, train both branches jointly.

Moreover, we can draw more conclusions by looking at Figure 6, which represents in a confusion matrix the percentage of hits and misses of classifying a mask of a concrete GT class as the predicted structure class. On the one hand, we can observe that the majority of the masks are being classified as the more frequent classes (skin, lesion, and pigment network). Hence, we could suspect that the balancing strategy for the structure classification loss is not adequate. On the other hand, we can also observe that the decoder is generating masks of all kinds of structures except for the last one (streaks), which is also the least frequent. Accordingly, we can conclude that the balancing method is working better for the segmentation task.

As it is shown in Figure 5, the decoder is producing lesion and skin masks quite decently. Nevertheless, the other masks are not similar to any of the other structures. This happens frequently in the results provided by our approach, which most of the times outputs accurate masks only for skin and lesion classes.

### 4.4 Interpretation: Sequences of gazes based on attention maps

Our approach provides explainability by means of a sequence of attention or gazes for each of the lesion images, for which it offers a diagnosis. These are obtained by computing the maximum of each of the $T$ attention maps computed by the ConvLSTM-based attention mechanism. The process is illustrated in Figure 7.

Additional examples of lesions together with their associated sequences of gazes, GT and predicted diagnosis, are shown in Figure 8, in order to perform a qualitative error analysis of our approach.

First, for the lesion A, we can appreciate that the network is focusing on the lesion and in the pigment network. The structure classification is not fine though, as it was mentioned before. As this structure only has pigment network apart from skin and lesion, the system has completed a meaningful diagnosis.

However, in lesion B, the network seems to focus on the blue stain and hair. This is a drawback of the attentive module, due to the fact that it helps the network to be focused on relevant areas. If the attention ends being focused at the wrong points, the final diagnosis could fail. Nevertheless, in this case, the diagnosis is correct.

In lesion C, four of the the predicted attentions match. However, the structure classes are not intuitive, since the first four instants are located in the lesion and, from the second timestep, the GT structure classes are *skin*. Nevertheless,

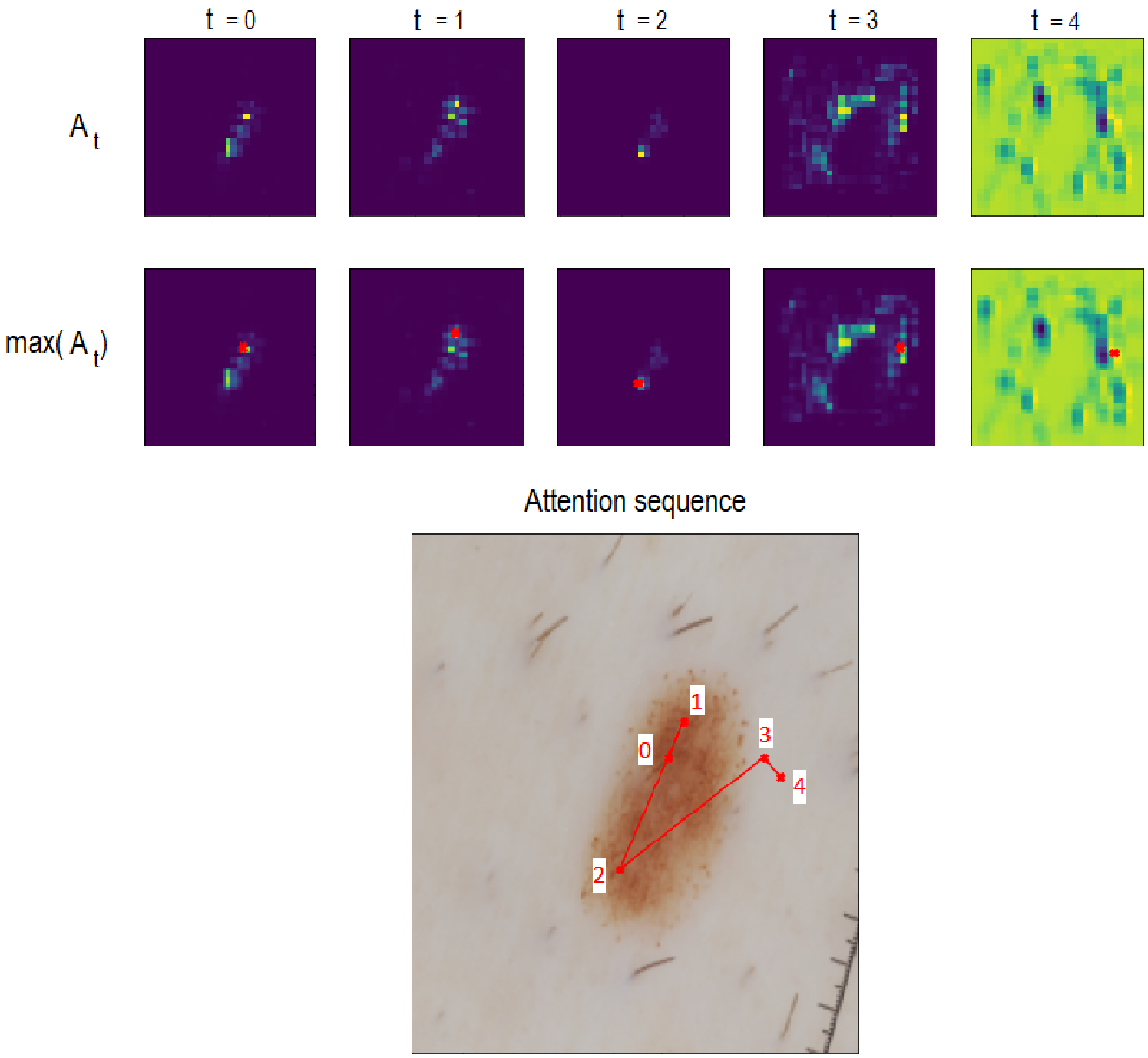

Fig. 7. Generation of sequences of gazes in the model proposed. Given the attention maps computed by the ConvLSTM-based attention mechanism (first row) for an example image taken from the ISIC 2017 [4] database (bottom left), the maximum for each map is computed (red dots in second row). Then, maximums are displayed over the original image (bottom right), indicating their order in the sequence.

these gazes are located near the border of the lesion and maybe they could be understood as the network evaluating the limit between the skin and the lesion.

Lastly, in lesion D, all the segmented structures are classified as skin, while the GT classes are skin, lesion and pigment network, which are all the structures of this lesion. Although the structure classification is not working properly, it seems that the segmentation captures the essence of these structures. Note that pigment network is the most frequent and widest structure within the database.

To sum up, we have seen through four examples the structures that the attention mechanism captures. Besides skin and lesion, only the pigment network structure is being recognized by the segmentation module. As we saw when analyzing the results on structure classification, the methods proposed for data balancing are not working as expected in the segmentation loss. Hence, this network could have performed better using other strategies for dealing with data imbalance and/or with a dermoscopic-balanced database.

## 5 Conclusions and Future Work

CAD systems based on Deep Learning have been proven to efficiently help to diagnose skin lesions. However, most of the proposed models in the literature are like black boxes for the dermatologist, given that they are not able to provide explainability about their results. Including attention mechanisms into the network architecture and/or combining the skin lesion classification task with other related ones in order to obtain extra relevant skin lesion information may make these systems trustworthy for medical experts.

In this paper, we have presented a system which provides an extended diagnosis based on sequences of gazes

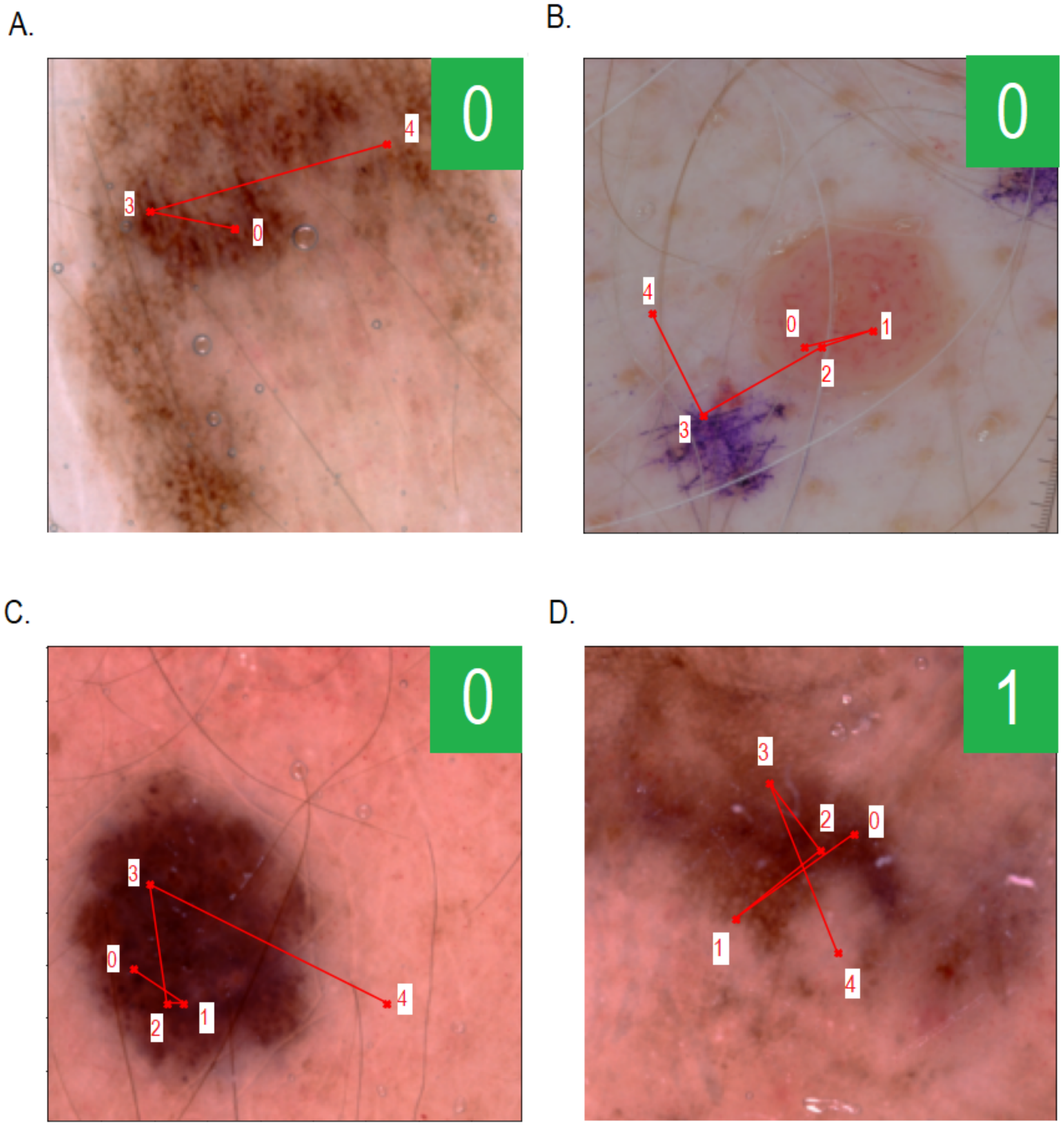

Fig. 8. Sequences of gazes or attention over example images taken from the ISIC 2017 [4] database. The number indicates the GT class (0 - benign, 1 - melanoma) and the color indicates if the prediction is right (green) or not (red). **A.** Predicted attention sequence: [0, 4]-skin. GT attention sequence: 0-pigment network, 1-skin, [2,3]-pigment network, 4-lesion. **B.** Predicted attention sequence: [0,1]-skin, 2-milia-like cysts, [3,4]-skin. GT attention sequence: 0-lesion, 1-skin, [2,3]-lesion, 4-skin. **C.** Predicted attention sequence: 0-lesion, [1,4]-skin. GT attention sequence: 0-lesion, 1-skin, 2-lesion, 3-skin, 4-lesion. **D.** Predicted attention sequence: [0,4]-skin. GT attention sequence: 0-pigment network, 1-skin, [2,4] -lesion.

in dermoscopic attributes. The benefit of these attention sequences is twofold: they are very intuitive and easy for the dermatologists and they provide explanability. The model proposed obtained decent results for the diagnosis task in comparison to other models in the literature and, in relation to the dermoscopic structures segmentation task, the model is learning how to segment skin and lesion reasonably well, but it is not able to recognize more specific structures, which are less frequent in lesions. However, to the best of our knowledge, the proposed model performs the first recurrent segmentation over skin lesion structures. Hence, we can conclude that this model is full of potential and, with further work on it, scores could be boosted.

Future work should include the evaluation of the attention sequences obtained. This could be done by comparing them with real GT sequences of eye fixations, which can be recorded by experts performing the skin lesion recognition task, using an eye tracker device. For the comparison of the scanpaths, the Jarodzka et al. measure [47] could be helpful. Concerning expert sequences of eye fixations, it would be also interesting to train the network with them, either using those in combination with the GT structure masks, or even as a substitute. This information would be very valuable and probably would significantly enhance the performance of the model.

Regarding the database used for our experiments, extending it by adding more images with a higher pixel-level GT content about dermoscopic structures should help to reduce the existing imbalance, at the same time it facilitates the learning stage of the system. Moreover, having less limited computational resources, as the ones provided by the Google Colab GPU environment, together with the consideration of more efficient strategies for hyperparameter tuning, would allow for a better search of the optimal configuration for the system.

In addition, evaluating different training strategies such as training in the first place the decoder branch for the recurrent segmentation of dermoscopic structures, which implies the most challenging task and, after a few epochs, start training the whole network jointly, could be also a good scheme to follow in future experiments.

Another strategy for the recurrent segmentation task could be considering a hierarchical approach. In an attempt to facilitate the learning of the least frequent dermoscopic features, the network would first segment and classify according to more generic structures (lesion/skin) to later identify more specific ones (pigment network, streaks, etc.).

Finally, trying other different balancing methods could also help acquiring better results, since the network is still not learning the most infrequent structures.

## References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020. DOI: 10.3322/caac.21590. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21590. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21590.

[2] A. S. of Cancer Oncology (ASCO, *Melanoma: Statistics*, urlhttps://www.cancer.net/cancer-types/melanoma/statistics, 2020.

[3] D. Gutman *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)*, 2016. arXiv: 1605.01397 `[cs.CV]`.

[4] N. C. F. Codella *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)*, 2018. arXiv: 1710.05006 `[cs.CV]`.

[5] N. Codella *et al.*, *Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)*, 2019. arXiv: 1902.03368 `[cs.CV]`.

[6] 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*, vol. 23, no. 3, pp. 1096–1109, 2018.

[7] I. Gonzalez-Diaz, "Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 547–559, 2018.

[8] Y. Yan, J. Kawahara, and G. Hamarneh, "Melanoma recognition via visual attention," in *International Conference on Information Processing in Medical Imaging*, Springer, 2019, pp. 793–804.

[9] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[10] A. Menegola *et al.*, "Recod titans at isic challenge 2017," *arXiv preprint arXiv:1703.04819*, 2017.

[11] 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*, 2016, pp. 770–778.

[12] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[15] 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*, 2016, pp. 770–778.

[16] J. Kawahara and G. Hamarneh, "Fully convolutional neural networks to detect clinical dermoscopic features," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 578–585, 2018.

[17] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, 2018.

[18] C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," *Pattern Recognition*, p. 107 413, 2020.

[19] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*, IEEE, 2013, pp. 273–278.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[21] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] S. Xingjian *et al.*, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[24] A. Salvador *et al.*, "Recurrent neural networks for semantic instance segmentation," *arXiv preprint arXiv:1712.00617*, 2017.

[25] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6688–6697.

[26] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 11–19. DOI: 10.3115/v1/P15-1002. [Online]. Available: https://www.aclweb.org/anthology/P15-1002.

[27] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[28] F. Wang *et al.*, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[29] Z. Yan, W. Liu, S. Wen, and Y. Yang, "Multi-label image classification by feature attention network," *IEEE Access*, vol. 7, pp. 98 005–98 013, 2019.

[30] Q. Guan and Y. Huang, "Multi-label chest x-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, vol. 130, pp. 259–266, 2020.

[31] A. Sinha and J. Dolz, "Multi-scale guided attention for medical image segmentation," *arXiv preprint arXiv:1906.02849*, 2019.

[32] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 370–378.

[33] Y. Yan, "Attention-based skin lesion recognition," Ph.D. dissertation, Applied Sciences: School of Computing Science, 2020.

[34] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018.

[35] 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*, 2016, pp. 770–778.

[36] 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.

[37] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[39] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3438–3446.

[40] G. Argenziano *et al.*, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Archives of dermatology*, vol. 134, no. 12, pp. 1563–1570, 1998.

[41] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[42] S. Bailey, *Step-by-step explanation of scoring metric*, urlhttps://www.kaggle.com/stkbailey/step-by-step-explanation-of-scoring-metric, 2018.

[43] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[44] R. Draelos, *The complete guide to auc and average precision: Simulations and visualizations*, urlhttps://glassboxmedicine.com/2020/07/14/the-complete-guide-to-auc-and-average-precision-simulations-and-visualizations/, 2020.

[45] G. developers, *Clasificación: Roc y auc*, urlhttps://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, 2020.

[46] 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.

[47] H. Jarodzka, K. Holmqvist, and M. Nyström, "A vector-based, multidimensional scanpath similarity measure," in *Proceedings of the 2010 symposium on eye-tracking research & applications*, 2010, pp. 211–218.