Expert Systems With Applications 118 (2019) 400–410

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An image-based segmentation recommender using crowdsourcing and transfer learning for skin lesion extraction

Amira Soudani[a], Walid Barhoumi[a,b,*]

[a] Université de Tunis El Manar, Institut Supérieur d'Informatique, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance (LIMTIC), 2 Rue Abou Rayhane Bayrouni, Ariana 2080, Tunisia
[b] Université de Carthage, Ecole Nationale d'Ingénieurs de Carthage, 45 Rue des Entrepreneurs, Tunis-Carthage 2035, Tunisia

ARTICLE INFO

Article history:
Received 3 May 2018
Revised 14 October 2018
Accepted 15 October 2018
Available online 15 October 2018

Keywords:
Segmentation recommender
Deep learning
Crowdsourcing
Transfer learning
VGG16
ResNet50

ABSTRACT

Deep learning is widely used in medical applications regarding the high performance it can achieve. In this paper, we propose a segmentation recommender based on crowdsourcing and transfer learning for skin lesion extraction. In fact, after collecting and pre-processing data from the ISIC2017 segmentation challenge, we tested two pre-trained architectures (VGG16 and ResNet50) to extract features from the convolutional parts. Then, a classifier with an output layer, composed of five nodes representing the segmentation methods' classes, was built. Thus, the proposed architecture is able to dynamically predict the most appropriate segmentation technique for the detection of skin lesions in any input image. Experimental results prove the capability of the proposed image-based method to improve the segmentation performance comparatively to the state of the art methods.

1. Introduction

Skin cancer is a major public health problem, with over five million newly diagnosed cases in the United States each year (Siegel, Miller, & Jemal, 2017). Invasive and in-situ malignant melanoma together form the most deadly skin cancer that comprises one of the most rapidly increasing cancers in the world (melanoma has an estimated incidence of 91,270 in the United States in 2018 (American Cancer Society, 2017)). In fact, according to the World Health Organization, about 132,000 new cases of melanoma are diagnosed worldwide each year, and invasive melanoma alone is responsible for 9,730 deaths each year. Fortunately, early diagnosis usually increases the chances for successful treatment. Indeed, melanoma is readily curable with simple excision, and the survival rate is generally very high, if it is recognized and effectively treated in its earliest stages. As a matter of fact, the first most common signs of melanoma are generally visible on an existing mole and it is possible to notice an increase in size, change in shape, crusting or bleeding. Depending on the behaviour of skin lesions, the melanoma can be recognized as benign or malignant. On the one hand, benign lesions show a more ordered and controlled growth. In most cases, they do not come back, and do not spread to other parts of the body. On the other hand, malignant lesions are made up of cells that generally grow out of control, and may invade other tissues of the body leading to death. Historically, the primary form of melanoma diagnosis had been unaided clinical examination. However, this kind of examination recorded commonly limited and variable accuracy, leading to significant challenges both in the early detection of disease and the minimization of avoidable biopsies and excisions of benign skin lesions that resemble to melanoma. Then, seen that pigmented lesions occur on the surface of the skin, there was a growing interest on visual inspection by specialists for early diagnosis of melanoma. Nevertheless, this manual inspection suffers from intra and inter-variability between experts, even for experienced ones (Barhoumi, Dhahbi, & Zagrouba, 2007), since early-stage melanomas may be difficult to distinguish from benign skin lesions with similar appearances. Indeed, although that dermoscopy has improved the diagnostic capability of trained clinicians, dermoscopy remains difficult to learn, and several studies have proved restrictions of dermoscopy when adequate training is not administered; and even with sufficient training, analysis remains highly subjective. Generally, the aforementioned tools of melanoma diagnosis are quite effective but take a plenty time and cause difficulties to patients and pathologists. Hence, there has been a growing need for computerized automatic analysis of skin lesion images providing promising directions for

[*] Corresponding author.
E-mail addresses: walid.barhoumi@enicarthage.rnu.tn, walid_barhoumi@yahoo.fr, walid.barhoumi@laposte.net (W. Barhoumi).

https://doi.org/10.1016/j.eswa.2018.10.029
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

computer-aided diagnosis (CAD) of melanoma. Therefore, over the last decade, there has been a significant rise in overall survival in patients with melanoma, and many studies have confirmed that this rise is mainly due to earlier diagnosis, notably using automated computer-assisted systems, which can be employed even on mobile devices (Rosado, Vasconcelos, Castro, & Tavares, 2015).

A typical CAD for automated melanoma diagnosis is performed in three different stages, namely lesion segmentation, feature extraction and classification (Filho, Ma, & Tavares, 2015; Ma & Tavares, 2017; Oliveira, Pereira, & Tavares, 2018a; Oliveira, Pereira, & Tavares, 2018b). The segmentation stage consists of extracting the lesion area. Then, specific features for each skin lesion are extracted in order to allow classifiers to decide whether it is cancerous or not-cancerous (Zortea, Flores, & Scharcanski, 2017). Given the strong impact of the segmentation quality on the overall precision of melanoma CAD, various works have been focused on developing accurate techniques for the automated segmentation of skin lesion images (Dey, Rajinikanth, Ashour, & Tavares, 2018; Ma & Tavares, 2016; Oliveira, Filho et al., 2016; Oliveira, Marranghello, Pereira, & Tavares, 2016). In fact, the importance of medical imaging of skin lesions for clinical decision-making has been steadily increasing over recent years, with greater emphasis on the design and implementation of automated segmentation techniques of these lesions (Oliveira, Papa, Pereira, & Tavares, 2018), what could be very beneficial in providing valuable assistance in the diagnosis and planning of the therapy. In this framework, with the ambition of improving melanoma diagnosis to reduce melanoma-related mortality, the International Skin Imaging Collaboration (ISIC) published in 2017, a large database that contains over 10,000 dermoscopic images, and a subset of these images has undergone annotation and markup by recognized skin cancer experts. Indeed, images in the ISIC archive have been derived from centres with expert pathology that can be deemed the gold standard. In particular, there was a reference standard expert-derived annotation for a subset of images aiming to serve as a public resource of images for the validation and the comparison of skin lesion segmentation techniques. More precisely, 600 images have been segmented against background normal skin and miscellaneous structures by expert dermatologists, and 21 segmentation techniques were applied on these images. This research database was a powerful resource of inspiration for us, since it contains both processed and unprocessed images, expert-derived ground-truth segmentations as well as a quantitative comparison of 21 relevant segmentation techniques on each image among the set of 600 skin lesion images. In fact, in the context of the lesion segmentation task of the ISIC2017 challenge,[1] participants were asked to submit automated predictions of lesion segmentation from 600 dermoscopic images in the form of binary masks (Codella et al., 2017), what permitted to collect 21 sets of prediction scores on each image of the final test set. The top ranked participant achieved an average Jaccard index score of 0.765, an accuracy of 93.4%, and a Dice coefficient of 0.849, using a variation of a fully convolutional network ensemble (Yuan, Chao, & Lo, 2017). However, our analysis of the global performance of the submitted segmentation techniques within the challenge, allows us to deduce that none technique was the best for the entire set of tested images. Furthermore, many least performing techniques, according to the mean value of the Jaccard score on the entire set of test images, are top ranked for some individual images, according to the same evaluation metric (Fig. 1). Our insight was inspired by this analysis that motivated us to study the possibility of selecting dynamically the adequate segmentation technique according to the input image. Indeed, the main goal of this work is to introduce an image-based mathematical model that defines a mapping of each relevant segmentation technique relatively to each class of lesion images. This model can be learned from the publicly available ISIC2017 challenge results, with the ambition of automatically recommending the adequate segmentation technique to any new skin lesion image. To do this, deep learning had been explored to exploit these large quantities of images, while adopting the principle of crowdsourcing, in order to recommend the most appropriate segmentation technique for the detection of skin lesions in any input image. Indeed, having a ground-truth benchmark relatively to the segmentation of melanoma (the ISBI dataset in our case), our objective is to select, in a dynamic manner, the most suitable technique for extracting the lesion zone for each image to be treated, without being obliged to make this choice in a global manner for a set of images. It is worth noting that although it was validated in the context of melanoma diagnosis, the proposed method can easily be adapted for the segmentation of any kind of medical images (Zagrouba & Barhoumi, 2002), provided that there is an annotated dataset and a set of segmentation techniques, that play the role of crowd experts, with their corresponding objective metrics.

The rest of this paper is organized as follows. In Section 2, we discuss the related work. Then, we describe in Section 3 the suggested method for automatically recommending the best technique of melanoma segmentation relatively to a given image. In order to prove the performance of the proposed method, extensive experiments are presented in Section 4. Finally, we produce a conclusion with some directions for future work in Section 5.

2. Related work

Most of works on lesion segmentation were based on the extraction of robust image features, what required pre-processing and post-processing steps (contrast enhancement, artifact removal, morphological operations...) in order to optimize the detection of lesions' borders (Barhoumi & Zagrouba, 2002). However, deep neural networks can automatically learn robust features from raw image data. Indeed, these networks proved their strength as a machine learning tool in various computer vision applications (Long, Shelhamer, & Darrell, 2015; Nam & Han, 2016; Zhu, Porikli, & Li, 2016) (object tracking, semantic segmentation...) and recently in medical image applications (Brosch et al., 2016; Fu, Xu, Wong, & Liu, 2016; Wang et al., 2015; Yang, Zhang, Guldner, Zhang, & Chen, 2016). In this section, we give a brief overview of the main deep learning techniques that have been used for object detection and classification, followed by techniques used for medical image segmentation, notably those validated in the context of skin lesion analysis towards melanoma diagnosis.

As part of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC2010), authors in Krizhevsky, Sutskever, and Hinton (2012) proposed a deep Conventional Neural Network (CNN, or ConvNet), the AlexNet, in order to classify 1.2 million high-resolution images into 1000 different classes. Their network was made up of five convolutional layers, some of which are followed by max-pooling layers, dropout layers and three fully connected layers. Besides, data augmentation and dropout techniques were applied in order to reduce overfitting. A variant of this model was presented in the ILSVRC2012 competition and it achieved a winning top-5 test error rate. In Simonyan and Zisserman (2014), authors from the Visual Geometry Group (VGG) of the university of Oxford proposed and evaluated a very deep CNN, the VGG Net, for large scale image classification. They adopted deep architecture (up to 19 layers) with very small convolution and max-pooling filters of size 3×3 and 2×2 throughout the whole network. They demonstrated that deeper networks provide better results, since their work was ranked, during the ILSVRC2014 challenge, as the first and the second best method in localization and classifica-

[1] https://challenge.kitware.com/#challenge/583f126bcad3a51cc66c8d9a

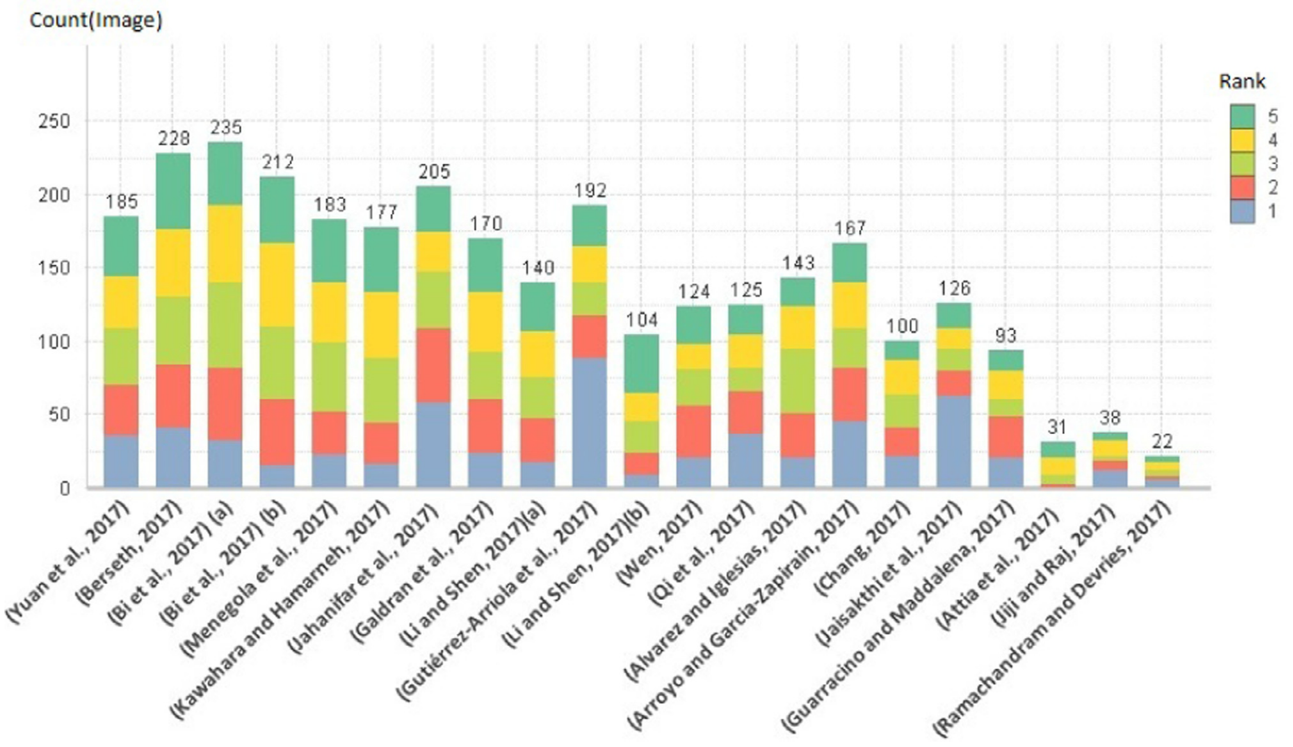


Fig. 1. Number of test images, for each segmentation method within the ISIC2017 challenge, for which the method is ranked among Top1, Top2, Top3, Top4 and Top5. We can deduce that all segmentation methods appear among Top5 for at least 22 images. The method of Gutiérrez-Arriola et al. (2017) is the best in Top1 since it records the best Jaccard index for 88 test images (among 600 ones), while the method of Attia et al. (2017) is not ranked first for any of the test images.

tion tasks, respectively. In He, Zhang, Ren, and Sun (2015), a deep residual network, the ResNet, with a depth up to 152 layers, was introduced. In fact, residual connections were added to the layers of the proposed network, and the layers were trained to fit the residual maps rather than the underlying maps. This architecture was evaluated on the ImageNet dataset (Russakovsky et al., 2014) and it won the first place in the ILSVRC2015 classification task. Furthermore, Ronneberger, Fischer, and Brox (2015) proposed a CNN, named U-Net, to deal with biomedical image segmentation. Indeed, an equal amount of up-sampling and down-sampling layers were combined, such that the up-sampling layers allow the network to propagate context information to higher resolution layers. Therefore, their network took into account the full context of the image and was able to be trained on small datasets. This model was extended in various works later. For instance, the 3D U-Net network was proposed in Çiçek, Abdulkadir, Lienkamp, Brox, and Ronneberger (2016) in order to segment a 3D volume from a sparse annotation. The network processed input data (3D volume) with 3D operations (3D convolution, 3D max-pooling…) and required some annotated 2D slices for training. Furthermore, Milletari, Navab, and Ahmadi (2016) inspired their 3D variant, named V-Net, from the U-Net architecture. It was based on volumetric fully convolutional neural network, while using an objective function based on Dice coefficient, in order to deal with 3D image segmentation. In Codella et al. (2015), deep learning was efficiently combined with sparse coding for feature extraction in conjunction with Support Vector Machines (SVM) for melanoma recognition. Indeed, a pre-trained CNN was used with an enhanced Bag-of-Visual-Words (BoVW) method based on sparse coding and average pooling over grayscale and colour images. A two-stage framework based on very deep CNNs was proposed for automated melanoma recognition (segmentation and classification). For the segmentation task, a Fully Convolutional Residual Network (FCRN), that incorporated a multi-scale contextual information integration scheme, was designed. The proposed FCRN was then integrated with a very Deep Residual Network (DRN) for the classification task in order to form a two-stage framework. Experiments have proved significant performance gains of the proposed framework. In fact, it won the ISIC2016, Skin Lesion Analysis Towards Melanoma Detection challenge at the International Symposium on Biomedical Imaging (ISBI) in 2016. Recently, the ISIC2017 challenge has been organized and several works have been evaluated for the melanoma segmentation task. In Yuan et al. (2017), authors proposed a framework based on fully Convolutional-Deconvolutional Neural Networks (CDNN) (Long, Shelhamer, & Darrell, 2014; Noh, Hong, & Han, 2015) to automatically segment skin lesions in dermoscopic images. Pixel-wise classification was performed and CDNN were essentially served as a filter that projects the entire input image to a map, where each element represents the probability that the corresponding input pixel belongs to the tumour class. In Berseth (2017), the U-Net architecture (Ronneberger et al., 2015) was employed to provide a probability estimate for each pixel in the original image and tried to include a Conditional Random Field (CRF) as a post-processing step. In Bi, Kim, Ahn, and Feng (2017), the deep Residual Networks (ResNets) (He et al., 2015) were exploited while following the Fully Convolutional Networks (FCN) architecture (Long et al., 2014). Convolutional and de-convolutional layers were added to up-sample the features maps derived from ResNet to output the score mask, and a multi-scale integration approach was used at the testing level to compute the final output. Authors in Menegola et al. (2017) proposed a network based on the work of Codella et al. (2015) while introducing modifications on their model, mainly on the loss function, the optimizer, the activation functions and the image pre-processing. In Kawahara and Hamarneh (2017), authors modified the VGG16 pre-trained over ImageNet to act as a pixel-wise classifier. They removed the fully connected layers, and converted the network such that the feature maps throughout the network were resized to match the size of the input before being concatenated. Moreover, a convolutional filter was added to the concatenated block to form the output for each of the four classes (Pigment Network, Negative Network, Milia-like Cysts, and Streaks). After a pre-processing step (hair removal and illumination correction), a supervised saliency map was constructed in Jahanifar, Tajeddin, Gooya, and Asl (2017) from the input image in order to extract an initial mask of the lesion. To that end, a random forest regressor, which takes a vector of regional image features as input, was trained in a multi-level manner. In particular, a dual-speed function was constructed based on the image gradient and the saliency map to evolve the initial contour toward the lesion boundary using a propagation model. In Almasni, Alantari, Choi, Han, and Kim (2018), authors proposed a segmentation method based on full resolution convolutional networks.

In this work, we introduce a new method for melanoma image segmentation based on crowdsourcing through segmentation results of the ISIC2017 challenge. We use best results based on the Jaccard index score on each image of the test dataset in order to collect needed informations to design a segmentation recommender based on deep neural networks. Unlike image segmentation frameworks that generally use a segmentation method based on the overall performance on the test dataset, despite it may provide poor results for several images separately, the proposed method allows to choose the more appropriate segmentation method we should apply regarding each input image. After selecting segmentation methods from the ISIC2017 challenge, we build our dataset that is mainly composed of the images of the test dataset provided by the challenge and the corresponding labels (best segmenter). Then, we train deep neural networks in order to learn how to predict, for any input dermoscopic image, the appropriate segmenter to apply. We evaluate the proposed method and we compare results with those provided by the best selected ISIC2017 segmentation methods, and the recorded experimental results demonstrate the performance gains of the suggested method.

3. Proposed method

According to the ISIC2017 challenge, best segmentation results can be given by different segmentation methods for dermoscopic images with high degree of visual similarity and sometimes for the same image. Likewise, the same segmentation method can perform the best results for images that look very different. Thereby, we ought to avoid working on original dermoscopic images which can deteriorate classification performance, especially that we have reduced the dataset since we should use only the test dataset (600 images) provided by the challenge to prepare our dataset. Thus, we propose a three-stage framework as shown in Fig. 2. First, we prepare the dataset by collecting, annotating, and applying basic augmentation operations in order to balance the set of images belonging to each label, before applying a resizing operation. Hence, we construct the ground-truth dataset that will serve to extract features based on transfer learning that can be performed using the pre-trained VGG16 architecture or ResNet50 architecture. The transfer learning is suited for classification task with small dataset, while reducing significantly the training time. In fact, we train a classifier using extracted features in order to make predictions. In this section, we firstly introduce the basics of the convolutional neural networks, then we specify how we prepared our dataset and the associated ground-truth labels and finally we describe the proposed image-based segmentation recommender.

3.1. Convolutional neural networks

Convolutional neural networks represent a class of deep and feed forward artificial neural networks. They are made up of a succession of layers composed of neurons that carry out different operations on input data. The aim of convolutional layers is to extract features from the input image. They take into account the 2D structure of the image and detect local features at different positions of the input feature maps. Down-sampling is performed by convolving the input feature maps with a fixed number of kernels with predefined sizes. The more convolutional layers we have, the more complex features will be extracted. Furthermore, each convolutional layer is always followed by a non-linear operation (*e.g.* ReLu, sigmoid…). In fact, non-linearity ensures that the model is more realist since the most of trained data would be non-linear. The pooling layers reduce the size of the feature maps by using various techniques such as max-pooling and average-pooling. For example, in the max-pooling technique, the highest value within extracted sub-windows is chosen based on predefined neighbourhood size. Therefore, the pooling layers make feature maps smaller and reduce the number of parameters and the complexity in the network, what particularly allows to control the overfitting. In addition, they make the network invariant to small transformations in the input image. Lastly, the fully connected layers are usually used at the end of the network. They represent regular layers of neurons in a neural network where all the neurons are connected to those of the previous layer so densely connected. The main goal of the fully connected layers is to use features provided by previous layers in order to classify the input image into various classes based on the training data.

3.2. Dataset and ground-truth generation

The dataset used in the ISIC2017 challenge is composed of dermoscopic images of size 64×64 pixels. It is divided into training (2000 images), validation (150 images) and test (600 images) datasets. In order to build our dataset, we used the 600 images of the challenge test dataset and we annotated each of them manually with the label *lab* of the segmentation method that recorded the best Jaccard index score for this image. In fact, the total number of final submitted works to this challenge was 21. These 21 methods were ranked based on the average Jaccard index score (equation 4) on the test dataset. The method proposed by Yuan et al. (2017) was ranked first in this challenge. We notice that it did not always provide the highest Jaccard index score separately for each image among the test dataset (Fig. 1). For example, for the images $ISIC\_12095$, $ISIC\_12134$ and $ISIC\_14792$, the method of Bi et al. (2017) provided the best scores. Furthermore, the method of Yuan et al. (2017) is recording a null score for the image $ISIC\_16034$. Thus, we can deduce that using one segmentation method may degrade performance, as it may not be adequate with one or more images of the test dataset. Thereby, the main idea of our work is to conceive a segmentation recommender based on the transfer learning and the crowdsourcing in order to decide the adequate segmentation method to be used for any dermoscopic image in input. For this purpose, we started by studying the rank of the submitted methods (Alvarez & Iglesias, 2017; Arroyo & Garcia-Zapirain, 2017; Attia, Hossny, Nahavandi, & Yazdabadi, 2017; Berseth, 2017; Bi et al., 2017; Chang, 2017; Galdran et al., 2017; Guarracino & Maddalena, 2017; Gutiérrez-Arriola, Gómez-Álvarez, Osma-Ruiz, Sáenz-Lechón, & Fraile, 2017; Jahanifar et al., 2017; Jaisakthi, Aravindan, & Mirunalini, 2017; Jiji & Raj, 2017; Kawahara & Hamarneh, 2017; Li & Shen, 2017; Menegola et al., 2017; Qi, Le, Li, & Zhou, 2017; Ramachandram & Devries, 2017; Wen, 2017; Yuan et al., 2017) based on the Jaccard index scores that are available on the challenge's leaderboard of the final segmentation submission. In Fig. 1, we showed for each method the number of images for which it was ranked as first, second, third, fourth or fifth best method of the challenge. This figure allows us to choose the methods that will be used to col-

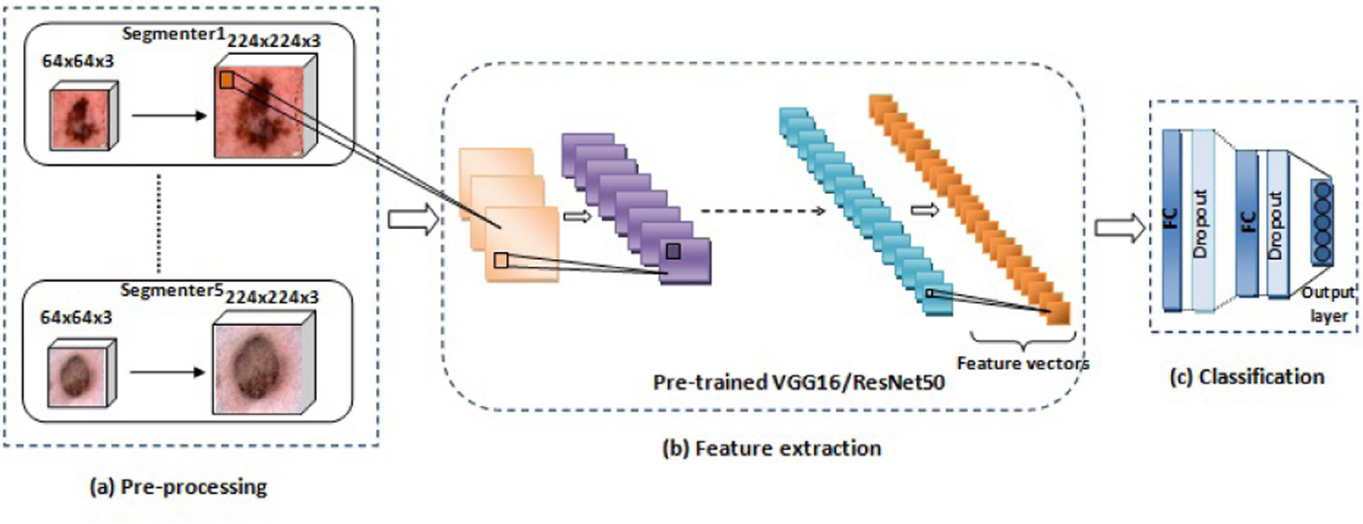


Fig. 2. Proposed melanoma segmentation recommender. (a) Pre-processing to prepare the dataset: images are first organized into sets that represent the classes' labels, and then data augmentation and resizing operations are applied. (b) Feature extraction based on transfer learning with pre-trained VGG16 or ResNet50 convolutional neural networks. (c) Classification is trained on the top of the feature extraction block: it is composed of two fully connected layers followed by dropout operations and an output layer with five output nodes that represent the classes of the selected segmentation methods.

Table 1
Ranking of the ISIC2017 methods according to the average of segmented test images classified in Top1, Top3 and Top5 (best values are in bold) .

Method	Top1	Top3	Top5	Average	Rank
Yuan et al. (2017)	5.83%	18.16%	6.16%	10.05 %	5
Berseth (2017)	6.83%	21.66%	7.6%	12.03 %	4
Bi et al. (2017)	5.33%	23.33%	**7.83%**	12.16 %	3
Bi et al. (2017)	2.5%	18.33%	7.06%	9.29 %	7
Menegola et al. (2017)	3.83%	16.5%	6.10%	8.81 %	8
Kawahara and Hamarneh (2017)	2.67%	14.66%	5.90%	7.74 %	12
Jahanifar et al. (2017)	9.67%	**24.5%**	6.83%	13.66 %	2
Galdran et al. (2017)	4%	15.33%	5.73%	8.35 %	9
Li and Shen (2017)	2.83%	12.5%	4.66%	6.66 %	14
Gutiérrez-Arriola et al. (2017)	**14.66%**	23.33%	6.4%	**14.79%**	1
Li and Shen (2017)	1.5 %	7.5%	3.46%	4.15 %	18
Wen (2017)	3.33%	13.5%	4.13%	6.98%	13
Qi et al. (2017)	6%	13.66%	4.16%	7.94%	10
Alvarez and Iglesias (2017)	3.33%	15.66%	4.76%	7.91%	11
Arroyo and Garcia-Zapirain (2017)	7.5%	3%	5.56%	5.35%	17
Chang (2017)	3.67%	10.5%	3.33%	5.83%	15
Jaisakthi et al. (2017)	10.33%	15.5%	4.2%	10.01%	6
Guarracino and Maddalena (2017)	3.33%	10%	3.1%	5.47%	16
Attia et al. (2017)	0%	1.5%	1.03%	0.84%	21
Jiji and Raj (2017)	2%	3.66%	1.26%	2.30%	19
Ramachandram and Devries (2017)	0.83%	2%	0.73%	1.18%	20

lect provided results for the training process, since we can not use all the 21 methods of the challenge. In fact, given that our idea is to crowdsource results from the test dataset which is very small (600 images), using all segmentation methods can lead to get sets of labelled classes (a class *lab* encompasses images manually annotated with the label *lab* of the segmentation method *lab* that records the best Jaccard index score for these images) with too few images (even empty sets). Data augmentation seems to be as a solution to this problem. Nevertheless, in our work we cannot apply all transformations because the segmentation results could significantly change. Thus, we opted to work on the best five methods (*i.e.* $lab \in \{1, 2, 3, 4, 5\}$) in order to not deteriorate the training performance. In Table 1, we compute the rank of all segmentation methods submitted to the challenge, on the entire set of 600 test images, based on the average of Top1, Top3 and Top5 score. According to this table, we choose to train our network with methods of Yuan et al. (2017), Berseth (2017), Bi et al. (2017), Jahanifar et al. (2017) and Gutiérrez-Arriola et al. (2017), that will play the role of the crowd. It is worth noting that the images that have more than one best segmenter were duplicated and added to our dataset. Thereafter, we obtain 611 images distributed as shown in Table 2. However, the obtained dataset is not balanced, what can lead to unexpected decrease in the classification performance. This is due to the fact that the skewed distribution of the class instances forces the classification algorithm to be biased toward ma-

Table 2
Image distribution according to the selected segmentation methods.

Label	Segmenter	Number of images	Percentage
1	Yuan et al. (2017)	119	19.48%
2	Berseth (2017)	113	18.49%
3	Bi et al. (2017)	100	16.37%
4	Jahanifar et al. (2017)	127	20.78%
5	Gutiérrez-Arriola et al. (2017)	152	24.88%

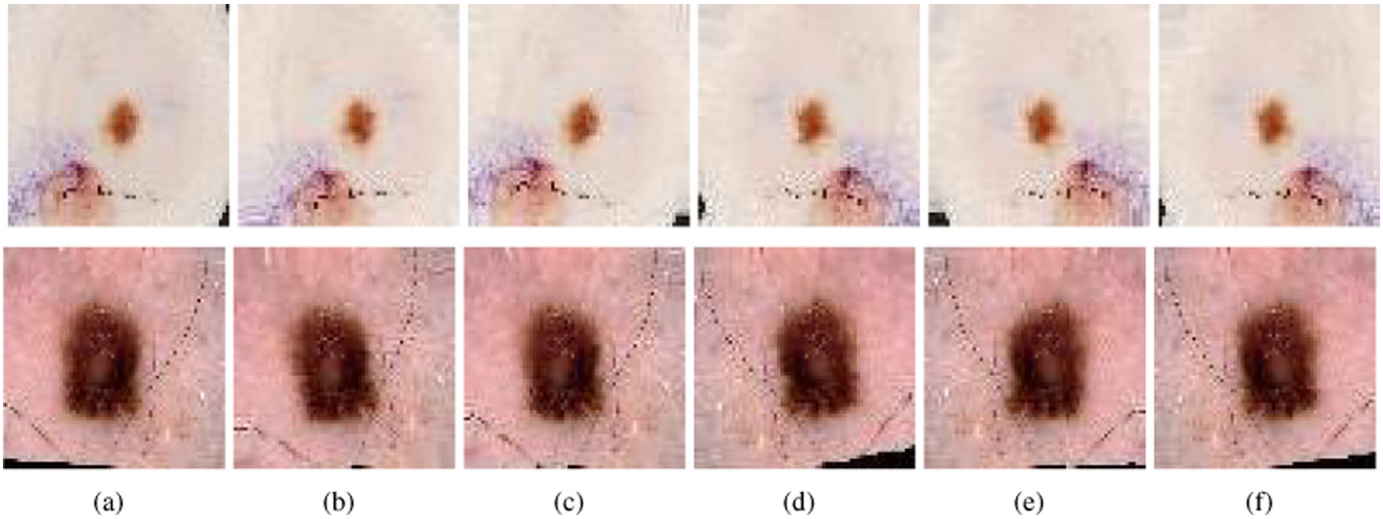


Fig. 3. Example of images generated by data augmentation. (a) Original images. (b-c) Results of applying random rotations of up to 10 degrees on the original images. (d-f) Results of applying horizontal flipping and random rotations of up to 10 degrees on the original images.

jority class. As a solution, we used data augmentation operations that will be presented in the pre-processing step (*c.f.* Section 3.3.1).

3.3. Proposed architecture: Segmentation recommender

The proposed architecture is composed of three major steps: the pre-processing step, the feature extraction step and finally the classification step. More details will be shown in the following paragraphs.

3.3.1. Pre-processing

In order to balance our dataset, we used augmentation techniques instead of manually balancing it through down-sampling which would further reduce its size. Thus, in order to obtain classes with equal numbers of samples, we chose basic augmentation techniques that preserve the appearance and the morphology of the original images (horizontal flipping and rotations of up to 10 degrees). Note that we opted for random rotations with small angles in order to keep the original morphology of the abnormality and to avoid shape changes due to common resizing methods. In Fig. 3, we show two original images and the corresponding augmentation results. For the ground-truth definition, we assign the original image's label to the augmented images. Thus, we obtain a dataset composed of 755 dermoscopic images. Fig. 4 illustrates a sample of images within the obtained dataset. This dataset was randomly divided into training data, validation data and test data that were distributed according to the corresponding segmenter. Therefore, we used five balanced sets of dermoscopic images where each one represents a class label. Moreover, each original image was resized to the size ($= 224 \times 224$ pixels) required by the pre-trained CNNs (VGG16 and ResNet50 in our case).

3.3.2. Feature extraction

For each dermoscopic image I of label lab, our goal is to extract image features $f = \Phi(I)$ that best discriminate it from other classes' labels. In fact, we performed transfer learning instead of training the deep neural network from scratch. We chose the VGG16 and the ResNet50 architectures since they achieved excellent results (He et al., 2015; Simonyan & Zisserman, 2014).

On the one hand, the VGG16 architecture is composed of five convolutional blocks with input map of size 224×224, and it uses small convolutional and max-pooling filters, of size 3×3 and 2×2, throughout the whole architecture. Each convolutional block consists of 2D convolution layers with spatial pooling through the max-pooling layer, and each convolutional layer is followed by a non-linear operation. Indeed, the rectified linear unit (*ReLu*) (1) aims to effectively mitigate the problem of gradient disappearance. Finally, the network ends with a classifier block composed of two fully connected layers with 4096 nodes each and an output layer with softmax activation function (2) and 1000 nodes.

$$f(x) = max(0, x), \tag{1}$$

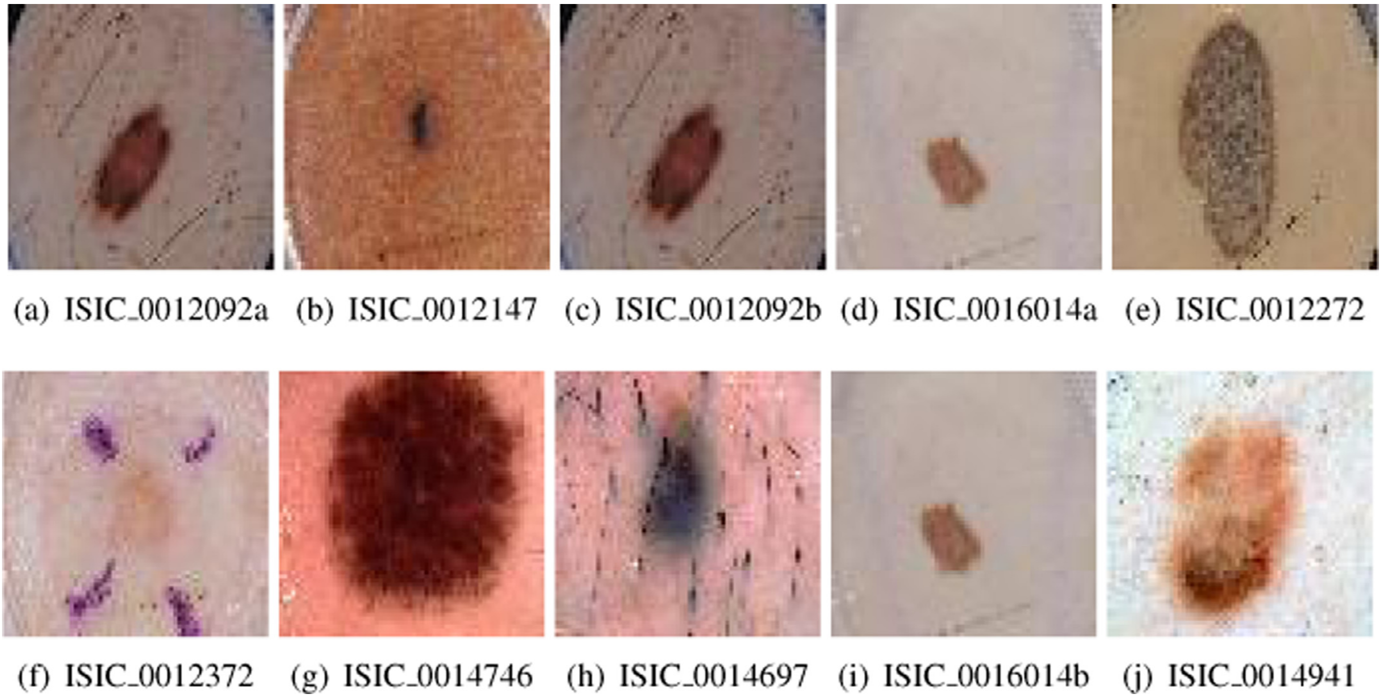
where x is an input unit.

$$y_j = \frac{exp(x_j)}{\sum_{k=1}^{K} exp(x_k)}, \quad with, \tag{2}$$

$$x_j = \sum_{k=1}^{M} h_k W_{k_j},$$

where y_j and x_j ($j \in \{1 \ldots K\}$) are respectively the output and the input units to the softmax layer, K is the total number of output nodes (classes), M denotes the number of input nodes, W_{k_j} denotes the weights connecting the penultimate layer x_j to softmax layer, and h_k is the activation of penultimate layer node.

On the other hand, the ResNet50 is a 50 layers Residual Network. It achieved very compromising results in the ImageNet and MS-COCO competitions (winner of the ILSVRC2015 classification task). This architecture is composed of 49 convolutional layers and one fully connected layer with input map of size 224×224. The convolutional part involves five blocks with small filters, of size 1×1 and 3×3. The first block is one convolutional layer followed by a max-pooling layer with a stride of 2. For the other blocks, each of them is composed of several residual blocks with stride of 2 for the last three blocks. The ResNet50 architecture ends with a fully connected layer with softmax activation function and 1000 nodes. The general idea of the transfer learning technique is to use models pre-trained on a large labelled dataset to fit a new task that can be totally different from the original one. Thus, the goal is to start learning process from patterns learned to solve different problem and so to generalize the model to a new dataset based on those patterns. The transfer learning is well adapted to small datasets and can be used as feature extractor. The main idea is to keep the convolutional parts and replace the fully connected layers with our classifier. In fact, we pre-trained the VGG16 and the ResNet50 on the ImageNet dataset (Russakovsky et al., 2014), which has been widely used to build various architectures given


(a) ISIC_0012092a (b) ISIC_0012147 (c) ISIC_0012092b (d) ISIC_0016014a (e) ISIC_0012272

(f) ISIC_0012372 (g) ISIC_0014746 (h) ISIC_0014697 (i) ISIC_0016014b (j) ISIC_0014941

Fig. 4. A sample of the training dataset. The pairs of images (a,b), (c,d), (e,f), (g,h), (i,j) belong to the segmenter of label 1, 2, 3, 4 and 5, respectively. We notice that some images were duplicated since two or more segmenters provided the same best Jaccard index score for those images (*e.g.* ISIC_0012092a and ISIC_0012092b, ISIC_0016014a and ISIC_0016014b).

Table 3
Hyperparameters' ranges.

Parameter	Values
Epochs	{20, 50, 100, 120}
Batch size	{20, 50, 80}
Learning rate (lr)	{0.001, 0.01, 0.1}
Dropout	{0, 0.1, 0.2, 0.3}
Loss function	$\{binary\ cross-entropy, categorical\ cross-entropy\}$
Optimizer	$\{Adam, SGD\}$
Momentum weight	[0.5,0.9]

its large size (1.2 million natural images) and we only used the convolutional part of these architectures. The activation maps were extracted before the last fully connected classifier of the networks. These maps are considered as the feature vectors that will be sent to the new classifier composed of dense layers, in order to train the features according to our dataset, and a softmax output layer with five outputs.

3.3.3. Classification

In order to classify our dataset, we trained a classifier on the top of the output of the last convolutional parts of each architecture (VGG16 and ResNet50). The classifier is composed of two fully connected (FC) layers that have 256 and 128 nodes, respectively, with ReLU activation function. Moreover, a dropout regularization is used after the FC layers in order to avoid overfitting, and thus to make the trained model better generalized. The output layer has five nodes that represent the output classes (the adequate segmenters) with softmax activation function. For the tuning of the models' hyperparameters, we performed a grid search (Domhan, Springenberg, & Hutter, 2015) with parameters ranges as detailed in Table 3. We tested various ratios of training, validation and test datasets, as shown in Table 4. The better validation score (= *validation_accuracy - validation_loss*) was given by the highest ratio of the training data. Thereby, we chose the hyperparameters that are recording the best validation score for the proposed models, with training data of ratio 80% from the whole dataset (Table 5). Therefore, we used a dropout regularization with rate of 0.2 for VGG16 and of 0.1 for ResNet50, and we adopted the binary cross-entropy loss function (3), which is minimizing the negative logarithmic likelihood between the prediction Y and the ground-truth data $\hat{Y}$.

$$L(Y,\hat{Y}) = -\frac{1}{K}\sum_{i=1}^{K}\left[\hat{y}_i log(y_i) + (1-\hat{y}_i) log(1-y_i)\right], \tag{3}$$

where, $\hat{y}_i$ is the ground-truth label of the ith training instance, y_i is the prediction for the ith training instance and K represents the number of output classes.

Furthermore, the Stochastic Gradient Descent (SGD) with momentum optimizer (4) allows the training of the networks with the same constant learning rate of 0.001 and momentum weight of 0.9.

$$v_t = \gamma v_{t-1} + \eta\nabla_\theta J(\theta), \quad with, \tag{4}$$

$$\theta \longleftarrow \theta - v_t,$$

where v_t is the current velocity vector, $J(\theta)$ is the objective function parameterized by a model's parameters θ, $\nabla_\theta J(.)$ represents the gradient of the objective function, η is the learning rate and γ denotes a fixed weight such that $\gamma \in [0, 1]$. The momentum helps accelerating the SGD in the relevant direction and mitigates oscillations. It performs this by adding a fraction of the update vector of the past time step to the current update vector. The loss function computes the error, between the output of the network and the ground-truth data, that will be back-propagated to the network in order to fit the weights of the layers. Once the training is achieved, we can predict for each image of the test dataset the label *lab* of the more appropriate segmenter.

Table 4
Impact of splitting training, validation and test datasets, using different ratios, on the global performance of the proposed architectures.

	Training	Validation	Test	Training accuracy	Validation accuracy	Training loss	Validation loss	Validation score
VGG16-based model	80%	10%	10%	0.912	0.803	0.199	0.479	0.323
	70%	20%	10%	0.801	0.799	0.495	0.499	0.300
	60%	30%	10%	0.895	0.791	0.295	0.514	0.277
ResNet50-based model	80%	10%	10%	0.817	0.809	0.414	0.446	0.362
	70%	20%	10%	0.871	0.806	0.296	0.477	0.328
	60%	30%	10%	0.885	0.806	0.282	0.479	0.326

Table 5
Top3 hyperparameters' combinations based on the validation accuracy given by the proposed architectures using VGG16 and ResNet50 (best combinations are in bold).

	Epochs	Batch size	lr	Dropout	Loss function	Optimizer	Momentum weight
VGG16-based model	**20**	**80**	**0.001**	**0.2**	**binary cross-entropy**	**SGD**	**0.9**
	20	50	0.001	0.3	binary cross-entropy	SGD	0.9
	50	50	0.001	0.3	binary cross-entropy	SGD	0.5
ResNet50-based model	**20**	**40**	**0.001**	**0.1**	**binary cross-entropy**	**SGD**	**0.9**
	50	20	0.01	0.2	binary cross-entropy	SGD	0.5
	100	20	0.01	0.1	binary cross-entropy	SGD	0.5

Table 6
Evaluation metrics.

Metric	Expression	Description
Accuracy	$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$	It is defined as the rate of correctly classified items.
Specificity	$Specificity = \frac{TN}{TN+FP}$	It refers to the rate of negative items rightly identified.
Sensitivity (Recall)	$Sensitivity = \frac{TP}{TP+FN}$	It is defined as the proportion of positive items correctly classified.
Precision	$Precision = \frac{TP}{TP+FP}$	It is the ratio of correctly predicted positive samples to the total predicted positive samples.
Jaccard index score	$Jaccard\_idx = \frac{TP}{TP+FN+FP}$	It measures similarity between two sample sets.
Dice coefficient	$Dice\_coef = \frac{2\times TP}{2\times TP+FN+FP}$	It is a statistical measure that is used for comparing the similarity of two sample sets.
F_1-score	$F_1 score = 2 * \frac{(Precision*Recall)}{(Precision+Recall)}$	The weighted average of Precision and Recall.
Average accuracy	$Avg\_accuracy = \frac{\sum_{i=1}^{K} \frac{TP_i+TN_i}{TP_i+FN_i+FP_i+TN_i}}{K}$	It expresses the overall average effectiveness of a classifier (K denotes the number of output classes).

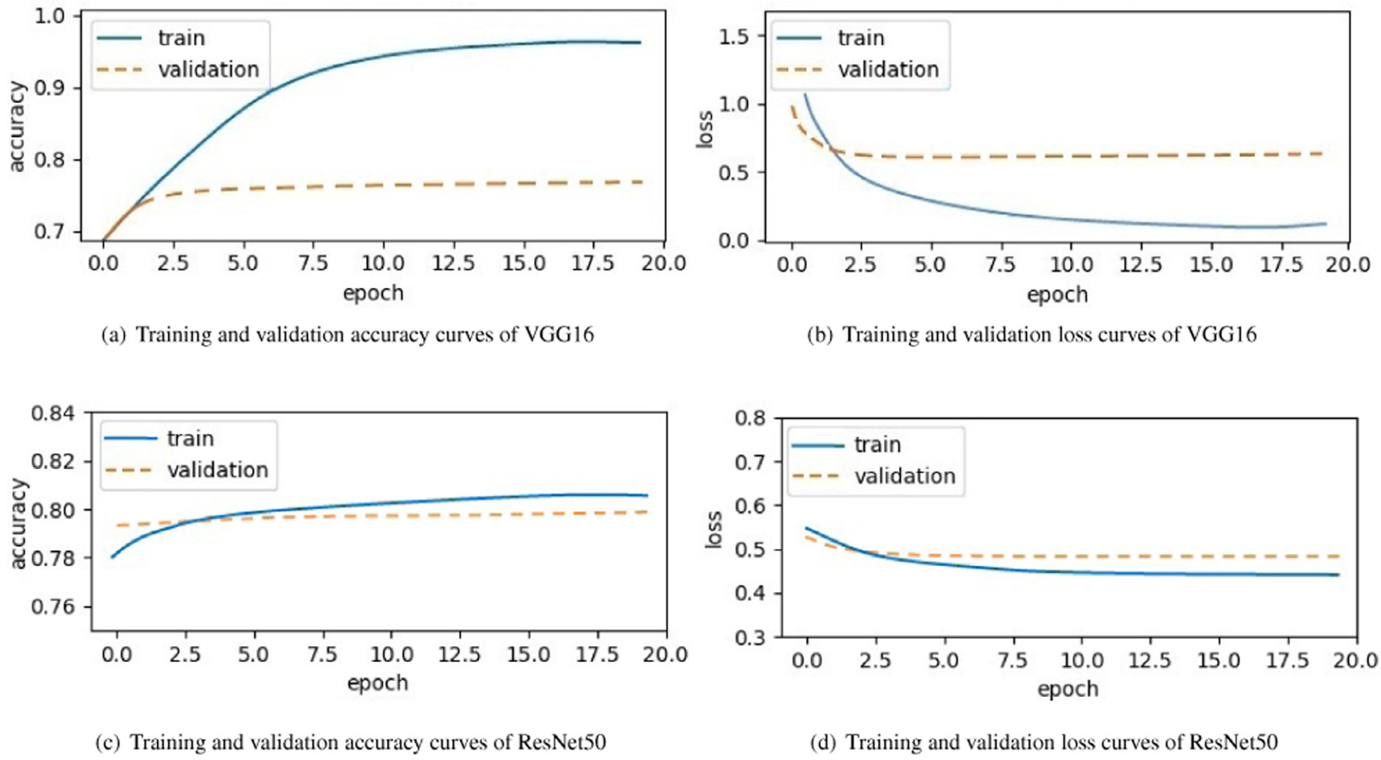
4. Experimental results

In order to evaluate the performance of the proposed method, we measured the following commonly used metrics: the accuracy, the specificity, the sensitivity, the Jaccard index score, the Dice coefficient as well as the F_1-score and the average accuracy, as reported in Table 6 (such that *TP* (True Positive) denotes the number of samples correctly classified as belonging to class *lab*, *FN* (False Negative) denotes the number of samples belonging to class *lab* but mis-classified, *TN* (True Negative) denotes the number of samples correctly classified as not belonging to class *lab*, and *FP* (False Positive) denotes the number of samples not belonging to class *lab* but mis-classified). The proposed method was implemented with Python on Keras library (Chollet et al., 2015) using a computer equipped with Intel Core i5 processor and 6 GB of RAM. The adopted models were trained on the training dataset (600 images ≃80%) with 20 epochs and a batch of size 80 and 40 for the first and second proposed models, respectively. In Fig. 5, we present the accuracy/loss curves of VGG16- and ResNet50-based models. Thus, we can deduce a high accuracy on the training dataset (above 0.97) comparatively to the validation dataset (about 0.73) (Fig. 5(a)) for VGG16 with loss curves that dip to reach very low score of 0.24 for the training dataset comparatively to 0.65 for the validation dataset (Fig. 5(b)). The training and validation accuracy provided by ResNet50 (Fig. 5(c)) are relatively high with close values while the loss curves reach low scores (Fig. 5(d)). The time needed for the training step of the proposed recommender is quite low ($\simeq 20s$) thanks to the used transfer learning.

In order to evaluate the proposed method against the state of the art segmentation methods that were selected from the challenge (Berseth, 2017; Bi et al., 2017; Gutiérrez-Arriola et al., 2017; Jahanifar et al., 2017; Yuan et al., 2017), we used the test dataset composed of 105 melanoma images. The accuracy of the classification task was evaluated by calculating the values of *TP, TN, FP* and *FN*. Those scores constitute the confusion matrix presented in Fig. 6, where we show in each cell the number of images predicted to belong to each class. The horizontal and vertical axis display the true and the predicted class labels, respectively. The first diagonal axis shows the rate of correctly classified images while all other entries represent misclassified items. We denote the best TP score given by the class 4 and a high correlation between several classes (*e.g.* (3,4) and (2,5)). This can be explained by the morphological similarity between images that belong to different classes. In Table 7, we report the most often used measures for classification assessment based on the values within the confusion matrix. The average accuracy of the proposed method is 0.763 and 0.732 for VGG16 and ResNet50, respectively. Lower sensitivity score is obtained for the third class ($lab = 3$) given its reduced number of *TP* with regards to the other classes. In fact, this is due to the huge number of augmented images that were generated in the pre-processing step in order to balance the dataset belonging to this class, although that we applied basic augmentation operations. Otherwise, the correlation between images belonging to different classes decreases the performance of the proposed segmentation recommender, regardless the used architecture (VGG16 or ResNet50).

In Table 8, we compared the performance of the proposed method against the five selected segmentation methods from the ISIC2017 challenge, while using the evaluation scores of the final test submission of the lesion segmentation task made publicly available.[2] Since our work was based on the Jaccard index score, we can see that we reach our goal by significantly improv-

[2] https://challenge.kitware.com/#phase/584b0afacad3a51cc66c8e24



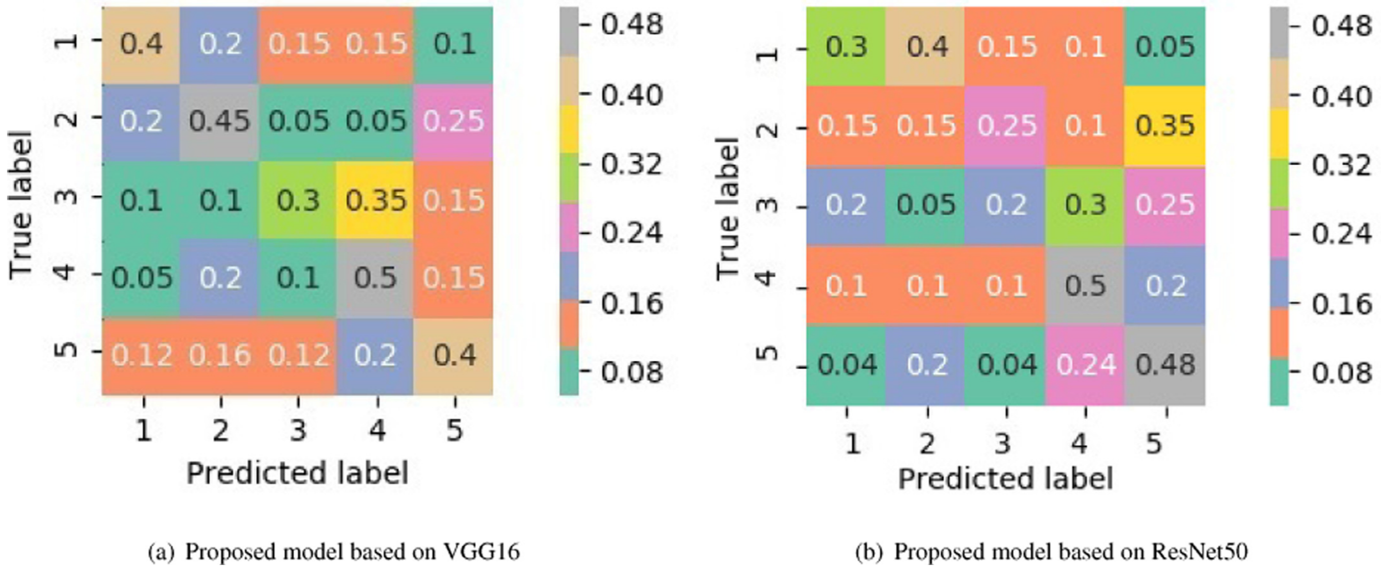
(a) Training and validation accuracy curves of VGG16

(b) Training and validation loss curves of VGG16

(c) Training and validation accuracy curves of ResNet50

(d) Training and validation loss curves of ResNet50

Fig. 5. Accuracy and loss curves of the proposed architectures (based on VGG16 and ResNet50). We reach high accuracy scores for both training and validation datasets during the training process, and the loss score is very low for the training dataset with VGG16.



(a) Proposed model based on VGG16

(b) Proposed model based on ResNet50

Fig. 6. Confusion matrix for all classes.

Table 7
Evaluation of the proposed models using the generated statistics for the five classes.

	Class	*Sensitivity*	*Specificity*	*Precision*	*Accuracy*	$\mathbf{F}_1$ score
VGG16-based model	1	0.4	0.882	0.444	0.790	0.421
	2	0.45	0.835	0.391	0.761	0.418
	3	0.3	0.984	0.4	0.780	0.342
	4	0.5	0.811	0.384	0.752	0.434
	5	0.4	0.837	0.434	0.733	0.416
ResNet50-based Model	1	0.3	0.882	0.375	0.771	0.333
	2	0.15	0.811	0.157	0.685	0.153
	3	0.2	0.870	0.266	0.742	0.228
	4	0.5	0.811	0.384	0.752	0.434
	5	0.48	0.887	0.413	0.714	0.444

Table 8
Evaluation of the proposed models comparatively to the selected ISIC2017 segmentation methods (best values are in bold).

Method	Jaccard_idx	Specificity	Sensitivity	Dice_coef	Accuracy
Yuan et al. (2017)	0.779	0.978	0.820	0.854	0.936
Berseth (2017)	0.766	0.981	0.815	0.842	0.929
Bi et al. (2017)	0.769	**0.990**	0.802	0.848	0.937
Jahanifar et al. (2017)	0.768	0.985	0.821	0.850	0.936
Gutiérrez-Arriola et al. (2017)	0.723	0.972	0.785	0.796	0.910
VGG16-based model	**0.786**	0.982	0.832	0.860	**0.938**
ResNet50-based model	**0.786**	0.983	**0.833**	**0.861**	0.937

ing this score on the used test dataset. Indeed, we obtain the best Jaccard index score ($= 0.786$) with the two proposed models (VGG16 and ResNet50) comparatively to the studied methods. Thus, the proposed method allows to predict the best segmentation method for an input image what leads to the improvement of the global Jaccard index score. Likewise, we record the best Dice coefficient and sensitivity scores ($= 0.861$ and $= 0.833$, respectively) for the ResNet50-based model and the best accuracy score ($= 0.938$) for the VGG16-based model. However, for the specificity metric, the proposed method records good scores and still better than (Yuan et al., 2017) (winner of the ISIC2017 Challenge). The slight decline of the proposed method performance, according to the specificity metric and comparatively to some studied methods, can be explained by the correlation between images that belong to different classes, what provokes the difficult discrimination between two different classes. Generally, experimental results demonstrate that predicting a segmentation technique, in a dynamic manner regarding each input image, could improve the performance, what leads to better skin lesion detection and extraction. Moreover, the proposed models based on VGG16 and ResNet50 show close results, what proves the efficiency of the transfer learning approach regardless the used architecture. Furthermore, we notice that transfer learning between two different tasks achieves good performance. In fact, the proposed models generalize well to a wide variety of classification problems, even if the predicted classes are not present in the ImageNet dataset.

5. Conclusion

In this work, we proposed a segmentation recommender based on crowdsourcing and deep learning. First, we built a new dataset using results of segmentation methods evaluated within the ISIC2017 challenge framework. Then, we performed transfer learning with VGG16 and ResNet50 architectures pre-trained on ImageNet dataset. We extracted feature vectors from the convolutional parts and we trained a classifier on the top in order to classify our test dataset of dermoscopic images into five classes that illustrate the most relevant segmentation methods. Experimental results showed that the proposed recommender records better Jaccard index score, Dice coefficient, sensitivity and accuracy, regarding the ISIC2017 selected segmentation methods. Thus, an image-based segmenter recommender that exploits deep learning seems to be a good idea to assist specialists to outperform convenient approaches that use a single segmentation method on the whole input images, regardless their appearance and the most appropriate segmentation method to be applied. Nevertheless, we was unable to use more than five classes due to the small size of the test dataset of the ISIC2017 challenge and the sensitivity of the segmented skin lesion images to data augmentation operations. Moreover, we was limited to 21 segmentation methods since our study is based on the Jaccard index score given by each image of the test dataset which is not always available outside the competition leaderboard. As perspectives, we aim to improve results while pre-processing the training set by decreasing the correlation between the image sets belonging to each class and by extracting more discriminative features. Furthermore, we plan to generalize this method on various applications under the constraint of data and ground-truth availability, ideally in the context of challenges (*e.g.* LiTS 2017 - Liver Tumour Segmentation Challenge, Lung CT Segmentation Challenge 2017...).

References

Almasni, M. A., Alantari, M. A., Choi, M., Han, S., & Kim, T. (2018). Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer Methods and Programs in Biomedicine, 162*, 221–231.

Alvarez, D., & Iglesias, M. (2017). k-means clustering and ensemble of regressions: An algorithm for the isic 2017 skin lesion segmentation challenge. *The Computing Research Repository (CoRR), abs/1702.07333*.

American Cancer Society (2017). Cancer facts & figures 2017 special section: Rare cancers in adults. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf.

Arroyo, J. L. G., & Garcia-Zapirain, B. (2017). Segmentation of skin lesions based on fuzzy classification of pixels and histogram thresholding. *The Computing Research Repository (CoRR), abs/1703.03888*.

Attia, M., Hossny, M., Nahavandi, S., & Yazdabadi, A. (2017). Spatially aware melanoma segmentation using hybrid deep learning techniques. *The Computing Research Repository (CoRR), abs/1702.07963*.

Barhoumi, W., Dhahbi, S., & Zagrouba, E. (2007). A collaborative system for pigmented skin lesions malignancy tracking. In *Ieee international workshop on imaging systems and techniques* (pp. 1–6).

Barhoumi, W., & Zagrouba, E. (2002). Boundaries detection based on polygonal approximation by genetic algorithms. In *Frontiers in artificial intelligence and applications* (pp. 1529–1533).

Berseth, M. (2017). ISIC 2017 - skin lesion analysis towards melanoma detection. *The Computing Research Repository (CoRR), abs/1703.00523*.

Bi, L., Kim, J., Ahn, E., & Feng, D. (2017). Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. *The Computing Research Repository (CoRR), abs/1703.04197*.

Brosch, T., Tang, L. Y. W., Yoo, Y., Li, D. K. B., Traboulsee, A., & Tam, R. (2016). Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging, 35*(5), 1229–1239. doi:10.1109/TMI.2016.2528821.

Chang, H. (2017). Skin cancer reorganization and classification with deep neural network. *The Computing Research Repository (CoRR), abs/1703.00534*.

Chollet, F. et al. (2015). Keras. https://keras.io.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3d u-net: Learning dense volumetric segmentation from sparse annotation. *The Computing Research Repository (CoRR), abs/1606.06650*.

Codella, N. C. F., Cai, J., Abedini, M., Garnavi, R., Halpern, A., & Smith, J. R. (2015). Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *6th international workshop on machine learning in medical imaging MLMI* (pp. 118–126). doi:10.1007/978-3-319-24888-2_15.

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., et al. (2017). 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). *The Computing Research Repository (CoRR), abs/1710.05006*.

Dey, N., Rajinikanth, V., Ashour, A. S., & Tavares, J. M. R. S. (2018). Social group optimization supported segmentation and evaluation of skin melanoma images. *Symmetry, 10*(2).

Domhan, T., Springenberg, J. T., & Hutter, F. (2015). Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *The 24th international conference on artificial intelligence IJCAI* (pp. 3460–3468).

Filho, M., Ma, Z., & Tavares, J. M. R. S. (2015). A review of the quantification and classification of pigmented skin lesions: From dedicated to hand-held devices. *Journal of Medical Systems, 39*(11), 177.

Fu, H., Xu, Y., Wong, D. W. K., & Liu, J. (2016). Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. In *Ieee 13th in-*

ternational symposium on biomedical imaging (ISBI) (pp. 698–701). doi:10.1109/ISBI.2016.7493362.

Galdran, A., Alvarez-Gila, A., Meyer, M. I., Saratxaga, C. L., Araujo, T., Garrote, E., et al. (2017). Data-driven color augmentation techniques for deep skin image analysis. *The Computing Research Repository (CoRR), abs/1703.03702*.

Guarracino, M. R., & Maddalena, L. (2017). Segmenting dermoscopic images. *The Computing Research Repository (CoRR), abs/1703.03186*.

Gutiérrez-Arriola, J. M., Gómez-Álvarez, M., Osma-Ruiz, V., Sáenz-Lechón, N., & Fraile, R. (2017). Skin lesion segmentation based on preprocessing, thresholding and neural networks. *The Computing Research Repository (CoRR), abs/1703.04845*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *The Computing Research Repository (CoRR), abs/1512.03385*.

Jahanifar, M., Tajeddin, N. Z., Gooya, A., & Asl, B. M. (2017). Segmentation of lesions in dermoscopy images using saliency map and contour propagation. *The Computing Research Repository (CoRR), abs/1703.00087*.

Jaisakthi, S. M., Aravindan, C., & Mirunalini, P. (2017). Automatic skin lesion segmentation using semi-supervised learning technique. *The Computing Research Repository (CoRR), abs/1703.04301*.

Jiji, G. W., & Raj, P. J. D. (2017). An extensive technique to detect and analyze melanoma: A challenge at the international symposium on biomedical imaging (ISBI) 2017. *The Computing Research Repository (CoRR), abs/1702.08717*.

Kawahara, J., & Hamarneh, G. (2017). Fully convolutional networks to detect clinical dermoscopic features. *The Computing Research Repository (CoRR), abs/1703.04559*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th international conference on neural information processing systems - volume 1*. In *NIPS'12* (pp. 1097–1105).

Li, Y., & Shen, L. (2017). Skin lesion analysis towards melanoma detection using deep learning network. *The Computing Research Repository (CoRR), abs/1703.00577*.

Long, J., Shelhamer, E., & Darrell, T. (2014). Fully convolutional networks for semantic segmentation. *The Computing Research Repository (CoRR), abs/1411.4038*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE conference on computer vision and pattern recognition (cvpr)* (pp. 3431–3440).

Ma, Z., & Tavares, J. M. R. (2017). Effective features to classify skin lesions in dermoscopic images. *Expert Systems with Applications, 84*, 92–101.

Ma, Z., & Tavares, J. M. R. S. (2016). A novel approach to segment skin lesions in dermoscopic images based on a deformable model. *IEEE Journal of Biomedical and Health Informatics, 20*(2), 615–623. doi:10.1109/JBHI.2015.2390032.

Menegola, A., Tavares, J., Fornaciali, M., Li, L. T., de Avila, S. E. F., & Valle, E. (2017). RECOD titans at ISIC challenge 2017. *The Computing Research Repository (CoRR), abs/1703.04819*.

Milletari, F., Navab, N., & Ahmadi, S. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *The Computing Research Repository (CoRR), abs/1606.04797*.

Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *IEEE conference on computer vision and pattern recognition (cvpr)* (pp. 4293–4302).

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. *The Computing Research Repository (CoRR), abs/1505.04366*.

Oliveira, R. B., Filho, M. E., Ma, Z., Papa, J. P., Pereira, A. S., & Tavares, J. M. R. (2016). Computational methods for the image segmentation of pigmented skin lesions: A review. *Computer Methods and Programs in Biomedicine, 131*, 127–141.

Oliveira, R. B., Marranghello, N., Pereira, A. S., & Tavares, J. M. R. (2016). A computational approach for detecting pigmented skin lesions in macroscopic images. *Expert Systems with Applications, 61*, 53–63. doi:10.1016/j.eswa.2016.05.017.

Oliveira, R. B., Papa, J. P., Pereira, A. S., & Tavares, J. M. R. S. (2018). Computational methods for pigmented skin lesion classification in images: review and future trends. *Neural Computing and Applications, 29*(3), 613–636.

Oliveira, R. B., Pereira, A. S., & Tavares, J. M. R. S. (2018a). Computational diagnosis of skin lesions from dermoscopic images using combined features. *Neural Computing and Applications*, 1–21.

Oliveira, R. B., Pereira, A. S., & Tavares, J. M. R. S. (2018b). Pattern recognition in macroscopic and dermoscopic images for skin lesion diagnosis. In *VipIMAGE 2017* (pp. 504–514). Cham: Springer International Publishing.

Qi, J., Le, M., Li, C., & Zhou, P. (2017). Global and local information based deep network for skin lesion segmentation. *The Computing Research Repository (CoRR), abs/1703.05467*.

Ramachandram, D., & Devries, T. (2017). Lesionseg: Semantic segmentation of skin lesions using deep convolutional neural network. *The Computing Research Repository (CoRR), abs/1703.03372*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *The Computing Research Repository (CoRR), abs/1505.04597*.

Rosado, L., Vasconcelos, M., Castro, R. N., & Tavares, J. (2015). From dermoscopy to mobile teledermatology. *Dermoscopy Image Analysis*, 385–418.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2014). Imagenet large scale visual recognition challenge. *The Computing Research Repository (CoRR), abs/1409.0575*.

Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians, 67*(1), 7–30.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *The Computing Research Repository (CoRR), abs/1409.1556*.

Wang, C., Yan, X., Smith, M., Kochhar, K., Rubin, M., Warren, S. M., et al. (2015). *Embc* (pp. 2415–2418). IEEE.

Wen, H. (2017). II-FCN for skin lesion analysis towards melanoma detection. *The Computing Research Repository (CoRR), abs/1702.08699*.

Yang, L., Zhang, Y., Guldner, I. H., Zhang, S., & Chen, D. Z. (2016). 3d segmentation of glial cells using fully convolutional networks and k-terminal cut. In *19th international conference on medical image computing and computer-assisted intervention MICCAI* (pp. 658–666).

Yuan, Y., Chao, M., & Lo, Y. (2017). Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. *The Computing Research Repository (CoRR), abs/1703.05165*.

Zagrouba, E., & Barhoumi, W. (2002). Semiautomatic detection of tumoral zone. *Image Analysis and Stereology, 21*, 13–18.

Zhu, G., Porikli, F., & Li, H. (2016). Robust visual tracking with deep convolutional neural network based object proposals on pets. In *IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 1265–1272). doi:10.1109/CVPRW.2016.160.

Zortea, M., Flores, E., & Scharcanski, J. (2017). A simple weighted thresholding method for the segmentation of pigmented skin lesions in macroscopic images. *Pattern Recognition, 64*, 92–104.