A novel approach for skin lesion symmetry classification with a deep learning model

Lidia Talavera-Martínez Pedro Bibiloni
Manuel González-Hidalgo
SCOPIA Research Group. University of the Balearic Islands, 07122 Palma, Spain
{l.talavera, p.bibiloni, manuel.gonzalez}@uib.es

Abstract

Symmetry is a distinguishing feature when diagnosing the malignancy of skin lesions, those with an irregular shape –asymmetry– are more likely to have a worse prognosis. This work presents a novel approach for skin lesion symmetry classification of dermoscopic images based on deep learning techniques. Also, we introduce a new dataset of labels for 615 skin lesions. During experimentation, we also evaluate whether it is beneficial to rely on transfer learning from pre-trained CNNs or traditional learning-based methods. As a result, we present a new simple, robust, and fast classification pipeline that outperforms methods based on traditional approaches or pre-trained networks, with a weighted-average F1-score of 64.5%.

1. Introduction

Malignant skin lesions, whose incidence rate is raising, poses a major problem for public health [7]. Although advanced cutaneous melanoma is still incurable, its early diagnosis can prevent malignancy, and increase the survival rate and treatment efficacy. The symmetry of a lesion is one of the predominant features when differentiating melanocytic patterns in skin lesions, since those with an irregular shape –asymmetry– are more likely to have a worse prognosis. However, the assessment of symmetry might be altered by the individual judgment of the observers, which depends on their experience and subjectivity [3]. Nowadays, specialists rely on the evaluation of dermoscopic images to complement their clinical analysis. These images enable the visualization of structures, shapes, and colors that are not discernible by a simple visual inspection. Their use has been shown to improve the diagnostic accuracy concerning simple clinical observation, being up to 10-30% more accurate [12]. To the best of our knowledge, only traditional computer vision approaches have been used to tackle this problem [1,6,11,15–18,21,22]. However, these approaches may hinder both the interpretability and the direct measurement of symmetry as an isolated feature, because symmetry features are usually integrated within a general-purpose system for the diagnosis of skin lesions, which usually relies on the automatic and simultaneous extraction of multiple features, such as color or texture [9, 14, 24]. Another difficulty is the lack of standard criteria on how to define the symmetry of the lesion —based on the shape, or the distribution of colors and textures— or on how to quantify it —using asymmetry indices or detecting the number of symmetry axes—, which complicates an objective comparison with other approaches.

In this work, we aim at going a step forward, implementing and assessing, for the first time, the adequacy of deep learning techniques to classify skin lesions according to their symmetry. To do so, 1) we propose a simple and robust CNN model that aims at classifying images depicting skin lesions into three classes, namely: "fully asymmetric", "symmetric with respect to one axis", or "symmetric with respect to two axes"; 2) we introduce the SymDerm dataset, a set of 615 labels for publicly available images according to the symmetry of skin lesions[1]; 3) we compare our results to other traditional methods, and 4) we also perform a transfer learning study where we evaluate whether it would be beneficial to use transfer learning from pre-trained CNNs or traditional learning-based methods.

2. SymDerm Dataset

Among the publicly available databases of skin lesions that we are aware of, only the PH^2 database [13] provides expert annotated data —200 dermoscopic images— regarding the lesion symmetry. To overcome the scarcity of data we consider two strategies. The first one is based on using images from the PH^2, which are already labeled, on which we have introduced realistic artifacts —simulated hair— following the approach described in [25]. That is,

[1] We will make the SymDerm dataset publicly accessible but available on demand.

we generate variations of the same lesions which are assigned the original symmetry annotation. With this task-consistent data augmentation technique, we enlarged the PH^2 dataset, from 200 to 438 annotated samples. The second strategy consisted of randomly extracting images from publicly available datasets, and asking three expert dermatologists to label according to their expertise the symmetry of the lesion as either “0: fully asymmetric”, “1: symmetric with respect to one axis”, or “2: symmetric with respect to two axes”. Thus, we managed to provide an additional set of 615 new expert annotations for the symmetry of skin lesions, which we called the SymDerm dataset. To solve the annotation discrepancy among the three experts, and to provide unified labels for the model training and metrics evaluation, we use the max voting method. It assigns the class that has the maximum number of votes among the experts. If there is a total disagreement because each expert selected a different label, we have decided to use the label of the expert with the most years of experience. It is worth mentioning that, in some lesions, dermatologists agree on the difficulty of ignoring the bias introduced by their clinical suspicion. Also, the annotation was affected by the limit that they set to determine what is symmetrical and what is not. In Figure 1, we show the confusion matrices between each expert labels and the max voting labels. As we can see, although dermatologists have tried to be as rigorous as possible, subjectivity causes some images to be interpreted differently between them. This is noticeable in the class of symmetric lesions with respect to one axis.

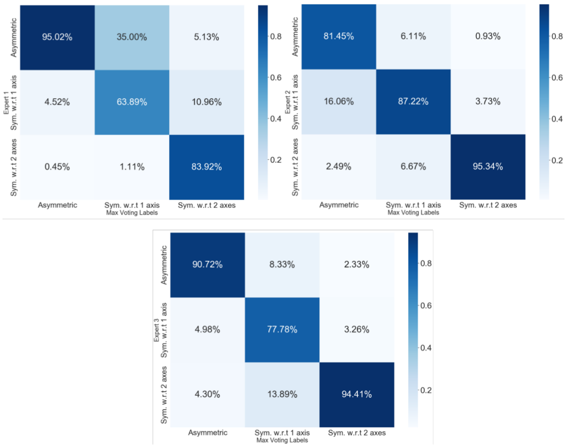


Figure 1. Confusion matrices between the labels of each expert and the maximum voting labels in the SymDerm dataset.

To sum up, we introduce a dataset with 1053 annotated images based on the symmetry of the lesion, which consists of 262 images from the EDRA2002 dataset [2], 438 images generated from the 200 samples of the PH^2 dataset, 177 images from the ISIC Data Archive [2], 41 images from the dermis dataset [3], and 135 from the dermquest dataset [4].

3. Experimental Framework

Architecture structure The CNN-based model we propose, which is detailed in Figure 2, is fed with images from the dataset of 1053 images depicted in Section 2, and it is composed of 10 layers. The first one, the input layer, resizes the images to a fixed size of 256 × 256 × 3, and it is followed by 3 blocks, each consisting of one 3 × 3 convolution layer and one down-sampling layer, which is applied by a two-stride 3 × 3 convolution to reduce the spatial resolution. In each of these blocks, we use 8, 16, and 32 filters, respectively. Then, we reduce the number of obtained high-level features from 32768 to 128, and later to 8 features, by means of two dense layers. Finally, we use a Softmax layer to output a 3-class classification of the symmetry of the lesions. The design of this model is based on the well-known VGG16 network [20], which has achieved very good results in numerous classification tasks. However, the VGG16 model is more complex due to a greater number of layers and therefore parameters. We believe that a smaller network would be more appropriate with the amount of data we have available for the problem in question.

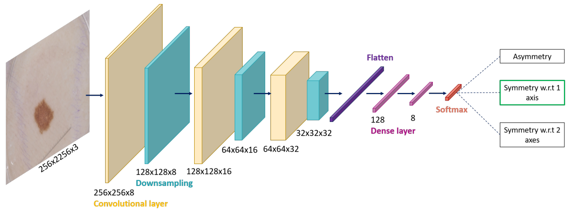


Figure 2. Architecture of our proposed network.

Transfer Learning Strategy It is achieved by 1) considering pre-trained CNNs on the ImageNet dataset [5], such as the VGG16 [20], the ResNet50 [8], and the Inception V3 [23] and further fine-tuned on our dataset to adapt the network to the problem at hand, and 2) using other samples annotated with a published method. Of the methods presented in the literature, our 3-class taxonomy only coincides with the work presented by Toureau *et al.* [26], and also based on the good results published by their algorithm, we use this method to label new data and used them together with the 1053 images we already had.

Implementation, learning details, and evaluation All the experiments were carried out with the Windows 10 Pro 64-bit operating system, a single NVIDIA Quadro P6000 24GB, and an HP Z640 Workstation with Intel®Xeon ®E5-2620 v4 @ 2,1 2133 8C. We implemented the proposed ar-

[2] www.isic-archive.com

[3] www.dermis.net

[4] Was deactivated on December 31, 2019

Table 1. Classification performance results averaged across the 5 fold of the cross-validation, for all possible setups of experiments, as well as the performance of the expert's agreement with the resulting max voting labels. The best results are in bold.

Exp				weighted-average		
		B.Acc	**Kappa score**	**Precision (Pr)**	**Recall (R)**	**F1-score**
0	Agreement Expert 1 and Max voting (3 classes)	0.803	0.778	0.869	0.865	0.865
1	Agreement Expert 2 and Max voting (3 classes)	0.900	0.852	0.921	0.907	0.911
2	Agreement Expert 3 and Max voting (3 classes)	0.838	0.790	0.875	0.870	0.869
3	Toureau *et al.* [26] (3 classes)	0.498 ± 0.018	0.291 ± 0.025	0.566 ± 0.015	0.562 ± 0.016	0.560 ± 0.015
4	Ali *et al.* [1] (3 classes)	0.479	0.256	0.563	0.523	0.539
5	**Proposed method (3 classes)**	**0.615 ± 0.019**	**0.429 ± 0.030**	**0.690 ± 0.026**	**0.627 ± 0.022**	**0.645 ± 0.021**
6	Stoecker *et al.* [21] (2 classes)	0.562	0.121	0.639	0.551	0.482
7	Ali *et al.* [1] (2 classes)	0.676	0.353	0.678	0.676	0.677
8	**Proposed method (2 classes)**	**0.719 ± 0.029**	**0.441 ± 0.059**	**0.735 ± 0.035**	**0.722 ± 0.029**	**0.718 ± 0.029**
9	VGG16 [20] pre-trained (3 classes)	0.584 ± 0.061	0.365 ± 0.084	0.683 ± 0.039	0.576 ± 0.061	0.597 ± 0.060
10	ResNet50 [8] pre-trained (3 classes)	0.341 ± 0.024	0.008 ± 0.021	0.123 ± 0.060	0.279 ± 0.104	0.143 ± 0.077
11	Inception V3 [23] pre-trained (3 classes)	0.436 ± 0.041	0.162 ± 0.064	0.558 ± 0.056	0.428 ± 0.044	0.420 ± 0.055
12	Proposed method + ExtraDataToureau (3 classes)	0.516 ± 0.062	0.301 ± 0.129	0.593 ± 0.094	0.544 ± 0.106	0.556 ± 0.111

chitecture using Keras [4] and trained it from scratch following a 5-fold stratified cross-validation strategy with 1) a batch size of 8, 2) randomly initialized weights, 3) the Adam [10] optimizer with a learning rate experimentally set to 10^{-4}, 4) the Weighted Categorical Cross-Entropy loss function following an early stopping policy based on monitoring the validation loss, and restoring the weights of the best epoch. We divided the dataset into 80% for training and 20% for the test and performed data augmentation in the training phase to improve the model's ability to generalize, restricted to those operations that do not distort the shape of the lesion, namely: flips in both horizontal and vertical directions.

To perform the quantitative evaluation, which provides an objective and comparable evaluation, we rely on several performance measures: Balanced Accuracy (B.Acc), Kappa score, Precision (Pr), Recall (R), and F_1-score. We also analyze the obtained results from a qualitative point of view, relying on the Gradient-Weighted Class Activation Maps (Grad-CAM) [19] to visualize which features of an input image contribute the most to activate the neurons to obtain the final decision of the model by calculating the gradient of the output with respect to the input image. With this tool, we provide better interpretability to increase the confidence of the specialists in the robustness of the model's decision-making that we present.

4. Results and Discussion

Our method was mainly conceived to classify the symmetry of the lesion into "fully asymmetric", "symmetric w.r.t one axis", and "symmetric w.r.t two axes". However, as far as we know, there is no method based on deep learning with which we can compare our proposed method. Hence, we have decided to assess its suitability against some methods present in the literature and based on traditional techniques. In this respect, we have decided to compare it with the version based on texture and shape of the method presented by Toureau *et al.* [26] and the method of Ali *et al.* [1], as they follow the same strategy when classifying the lesion symmetry according to 0, 1, or 2 perpendicular symmetry axes. Given the specifications of some other methods presented in the literature, as is the case of Stoecker *et al.*'s algorithm, we had to simplify our problem and consider it as a binary classification one. In this case, we merge classes "symmetric w.r.t one axis", and "symmetric w.r.t two axes" into the class "symmetric".

We present, in Table 1, the quantitative results of the experiments that have been carried out in this study. The first three rows —experiments 0 to 2—, refer to the agreement between the labeling of the test set established by each of the labels of each expert and the maximum voting labels. We can observe, from the results of experiments 3 to 5, that when evaluating a three-class classification problem, our proposed method with a B.Acc of 61.5%, a Kappa Score of 0.429, a weighted-average Pr of 69%, a weighted-average R of 62.7% and a weighted-average F1-score of 64.5%, widely overcomes the method of Toureau *et al.* and Ali *et al.*, with a minimum difference with respect to the first of 6.5% in all performance measures. On the other hand, when considering our problem as a binary one,—experiments 6 to 8—, we see how our proposed method with a B.Acc of 71.9%, a Kappa Score of 0.441, a weighted-average Pr of 73.5%, a weighted-average R of 72.2% and a weighted-average F1-score of 71.8%, substantially improves the results obtained by the method of Stoecker *et al.* and Ali *et al.* In both cases, the performance of all the measures are correlated, which indicates that we can trust them.

Next, we analyze the use of transfer learning to classify the symmetry of skin lesions using pre-trained networks, —experiments 9 to 11—. We conclude that the proposed method, which is a simpler network, achieves better results than other well-known and more elaborated networks that have been trained in a large database such as ImageNet. It should be noted that among the pre-trained networks, the VGG16 is the one that obtains the best results, which supports our premise that for the task in question, the most appropriate was a network with fewer parameters. Finally, with experiment 12, we show the results of introducing more data labeled by the method of Toureau *et al.* Although we expected that the results of our model with the extra data could exceed those of the baseline, this has not been the case, which makes sense after seeing the performance of the Toureau *et al.* method on our database.

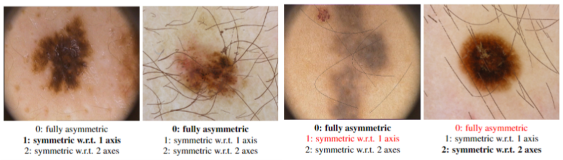

0: fully asymmetric
1: symmetric w.r.t. 1 axis
2: symmetric w.r.t. 2 axes
0: fully asymmetric
1: symmetric w.r.t. 1 axis
2: symmetric w.r.t. 2 axes
0: fully asymmetric
1: symmetric w.r.t. 1 axis
2: symmetric w.r.t. 2 axes
0: fully asymmetric
1: symmetric w.r.t. 1 axis
2: symmetric w.r.t. 2 axes

Figure 3. Example of correct (bold) and misclassified predictions (red) of our proposed model compared to their GTs (bold).

From a qualitative point of view, in Figure 3, we can see some examples of correct and misclassified predictions of our proposed model, respectively, when considering the three-class classification problem. Whereas, from Figure 4a, we can see in more detail how the model is able to correctly classify 85.29% of samples from class ‘Asymmetric’, 31.15% of samples from class ‘Symmetric w.r.t 1 axis’, and 65.88% of samples from class ‘Symmetric w.r.t 2 axes’. In terms of prediction errors, we can see that ‘Asymmetric’ samples are misclassified as any of the other two classes tend to occur with the same frequency, while for the classes ‘Symmetric w.r.t 1 axis’ and ‘Symmetric w.r.t 2 axes’, the errors occur more toward the class ‘Asymmetric’. It should be noted that most errors occur where there is more discrepancy on the part of the experts when it comes to labeling the images, that is in the class ‘Symmetric w.r.t 1 axis’. This last issue, as can be seen in Figure 4b, can be solved if we reduce the complexity of the problem to a taxonomy of two classes —‘Asymmetric’ vs. ‘Symmetric’—. To further study and understand the behavior of the network we show in Figure 5 the Grad-CAM of both correctly- and wrongly-classified samples. We can see that when classifying correctly a lesion, our proposed method focuses clearly both on the interior of the lesion and on its border with the skin region. This is not the case when considering misclassified samples. Finally, we notice that the black frame attracts great attention to the network, which can lead to errors.

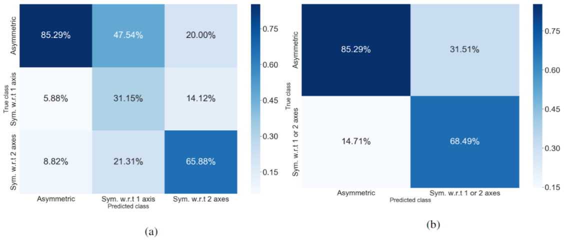
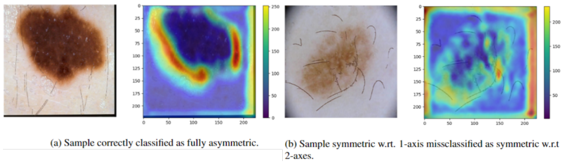


Figure 4. Confusion matrices resulting from (a) the base experiment of the proposed method, and (b) results from the proposed method when considering 2 classes: ‘Asymmetric’ and ‘Symmetric w.r.t 1 or 2 axes’.



(a) Sample correctly classified as fully asymmetric.

(b) Sample symmetric w.rt. 1-axis missclassified as symmetric w.r.t 2-axes.

Figure 5. Example of extracted Grad-CAM from the network’s last convolutional layer (right column) for two images (left column).

5. Conclusions

In this work, we have presented a novel CNN-based method for the task of skin lesion symmetry classification on dermoscopic images. Also, we have introduced a set of 615 labels for publicly available images according to the symmetry of skin lesions, which have been labeled by expert dermatologists. From the results, we conclude that our proposed method widely overcomes the traditional methods of Toureau *et al.* Stoecker *et al.*, and Ali *et al.* in all performance measures. These good results demonstrate the convenience of CNNs for the task at hand. However, the performance of our proposed method for the three-class problem is limited by the experts’ discrepancy. Another strength of this work compared to most traditional methods is that it does not require segmentation of the lesion to obtain symmetry of the lesion. Regarding the transfer learning approach, we observe how shallow networks obtain better results in the classification, mainly due to the scarcity of data, and the how our model has been able to discard low-quality information, when new data was incorporated, preventing its performance from collapsing. However, we conclude that it is preferable not to use this low-quality information. Finally, as future work, we aim to study more in-depth, some aspects of our work such as the number of images used to train the network and remove the black frames in the images.

Acknowledgements

This paper is part of the R+D+i Project PID2020-113870GB-I00, funded by MCIN/AEI/10.13039/501100011033/. One of the authors also benefited from the fellowship BES-2017-081264 conceded by the *Ministerio de Economía, Industria y Competitividad*.

References

[1] Abder-Rahman Ali, Jingpeng Li, and Sally Jane O'Shea. Towards the automatic detection of skin lesion shape asymmetry, color variegation and diameter in dermoscopic images. *Plos one*, 15(6):e0234352, 2020. 1, 3

[2] Giuseppe Argenziano, HP Soyer, V De Giorgi, Domenico Piccolo, Paolo Carli, and Mario Delfino. *Interactive atlas of dermoscopy (Book and CD-ROM)*. EDRA Medical Publishing & New media, 2000. 2

[3] Giuseppe Argenziano, H Peter Soyer, Sergio Chimenti, Renato Talamini, Rosamaria Corona, Francesco Sera, Michael Binder, Lorenzo Cerroni, Gaetano De Rosa, Gerardo Ferrara, et al. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet. *Journal of the American Academy of Dermatology*, 48(5):679–693, 2003. 1

[4] François Chollet et al. Keras: The python deep learning library. *Astrophysics Source Code Library*, 2018. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] D Gutkowicz-Krusin, M Elbaum, P Szwaykowski, and AW Kopf. Can early malignant melanoma be differentiated from atypical melanocytic nevus by in vivo techniques? part ii. automatic machine vision classification. *Skin Research and Technology*, 3(1):15–22, 1997. 1

[7] Gery P Guy Jr, Cheryll C Thomas, Trevor Thompson, Meg Watson, Greta M Massetti, and Lisa C Richardson. Vital signs: melanoma incidence and mortality trends and projections—united states, 1982–2030. *MMWR. Morbidity and mortality weekly report*, 64(21):591, 2015. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[9] Hitoshi Iyatomi, Hiroshi Oka, M Emre Celebi, Masahiro Hashimoto, Masafumi Hagiwara, Masaru Tanaka, and Koichi Ogawa. An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. *Computerized Medical Imaging and Graphics*, 32(7):566–579, 2008. 1

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint:1412.6980*, 2014. 3

[11] HF Lorentzen, K Weismann, and F Grønhøj Larsen. Structural asymmetry as a dermatoscopic indicator of malignant melanoma–a latent class analysis of sensitivity and classification errors. *Melanoma research*, 11(5):495–501, 2001. 1

[12] Justine Mayer et al. Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. *Medical journal of Australia*, 167(4):206–210, 1997. 1

[13] TF Mendonça, ME Celebi, T Mendonça, and JS Marques. PH^2: A public database for the analysis of dermoscopic images. *Dermoscopy Image Analysis*, 2015. 1

[14] Pietro Rubegni, Gabriele Cevenini, Marco Burroni, Roberto Perotti, Giordana Dell'Eva, Paolo Sbano, Clelia Miracco, Pietro Luzi, Piero Tosi, Paolo Barbini, et al. Automated diagnosis of pigmented skin lesions. *International Journal of Cancer*, 101(6):576–580, 2002. 1

[15] Margarida Ruela, Catarina Barata, Jorge S Marques, and Jorge Rozeira. A system for the detection of melanomas in dermoscopy images using shape and symmetry features. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 5(2):127–137, 2017. 1

[16] Philippe Schmid-Saugeona, Joël Guillodb, and Jean-Philippe Thirana. Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics*, 27(1):65–78, 2003. 1

[17] Stefania Seidenari, Giovanni Pellacani, and Costantino Grana. Pigment distribution in melanocytic lesion images: a digital parameter to be employed for computer-aided diagnosis. *Skin Research and Technology*, 11(4):236–241, 2005. 1

[18] Stefania Seidenari, Giovanni Pellacani, and Costantino Grana. Early detection of melanoma by image analysis. In *Bioengineering of the Skin*, pages 331–338. CRC Press, 2006. 1

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3

[21] William V Stoecker, William Weiling Li, and Randy H Moss. Automatic detection of asymmetry in skin tumors. *Computerized Medical Imaging and Graphics*, 16(3):191–197, 1992. 1, 3

[22] W Stolz. ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. *Eur. J. Dermatol.*, 4:521–527, 1994. 1

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2, 3

[24] Lidia Talavera-Martinez, Pedro Bibiloni, and Manuel González-Hidalgo. Computational texture features of dermoscopic images and their link to the descriptive terminology: A survey. *Computer methods and programs in biomedicine*, 182:105049, 2019. 1

[25] L. Talavera-Martínez, P. Bibiloni, and M. González-Hidalgo. Hair segmentation and removal in dermoscopic images using deep learning. *IEEE Access*, 9:2694–2704, 2021. 1

[26] Vincent Toureau, Pedro Bibiloni, Lidia Talavera-Martínez, and Manuel González-Hidalgo. Automatic detection of symmetry in dermoscopic images based on shape and texture. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 625–636. Springer, 2020. 2, 3